

Distributed Information-Theoretic Biclustering

Georg Pichler, *Student Member, IEEE*, Pablo Piantanida,
Member, IEEE, and Gerald Matz, *Senior Member, IEEE*

Abstract

We study a novel multi-terminal source coding setup motivated by the biclustering problem. Two separate encoders observe two stationary, memoryless sources X^n and Z^n , respectively. The goal is to find rate-limited encodings $f(x^n)$ and $g(z^n)$ that maximize the mutual information $I(f(X^n); g(Z^n))/n$. We present non-trivial outer and inner bounds on the achievable region for this problem. These bounds are also generalized to an arbitrary collection of stationary, memoryless sources. The considered problem is intimately connected to distributed hypothesis testing against independence under communication constraints, and hence our results are expected to apply to that setting as well.

I. INTRODUCTION

The recent decades witnessed a rapid proliferation of data available in digital form in a myriad of repositories such as internet fora, blogs, web applications, news, emails and the social media bandwagon. A significant part of this data is unstructured and it is thus hard to extract the relevant information. This results in a growing need for a fundamental understanding and efficient methods for analyzing data and discovering valuable and relevant knowledge from it in the form of structured information.

When specifying certain hidden (unobserved) features of interest, the problem then consists of extracting those relevant features from a measurement, while neglecting other, irrelevant features. Formulating this idea in terms of lossy source compression [1], we can quantify the complexity of the encoded data via its rate and the quality via the information provided about specific (unobserved) features.

In this paper, we introduce and study the distributed biclustering problem from a formal information-theoretic perspective. Given correlated samples X_1, X_2, \dots, X_K observed at different encoders, the aim is to extract a description from each sample, such that the descriptions are maximally informative about each other. In other words, the k -th encoder tries to find a (lossy) description U_k of its observation X_k subject to complexity requirements (coding rate), such that the mutual information between two disjoint subsets of descriptions $(U_k)_{k \in \mathcal{A}}$ and $(U_k)_{k \in \mathcal{B}}$ is maximized. The goal is to characterize the optimal tradeoff between *relevance* (mutual information between the descriptions) and *complexity* (encoding rate).

A. Distributed Biclustering

As a clustering technique, *biclustering* (or *co-clustering*) was first explicitly considered by Hartigan [2] in 1972. A historical overview of biclustering including additional background can be found in [3, Section 3.2.4]. In general, given an $N \times M$ data matrix (a_{nm}) , the goal of a biclustering algorithm [4] is to find partitions $B_k \subseteq \{1, \dots, N\}$ and $C_l \subseteq \{1, \dots, M\}$, $k = 1 \dots K$, $l = 1 \dots L$ such that all the “biclusters” $(a_{nm})_{n \in B_k, m \in C_l}$ are in a certain sense homogeneous. The measure of homogeneity of the biclusters depends on the specific application. The method received renewed attention when Cheng and Church [5] applied it to gene expression data. Many biclustering algorithms have been developed since in this field (e.g., see [6] and references therein). An introductory overview of clustering algorithms for gene expression data can be found in the lecture notes [7]. The information bottleneck method, which can be viewed as a uni-directional information-theoretic variant of biclustering, was successfully applied to gene expression data as well [8].

The material in this paper was in part presented at the 53rd Annual Allerton Conference on Communications, Control, and Computing, 2015. Part of this work was also submitted to the 2016 IEEE International Symposium on Information Theory.

G. Pichler and G. Matz are with the Institute of Telecommunications, TU Wien, Vienna, Austria. P. Piantanida is with Laboratoire des Signaux et Systèmes (L2S, UMR8506), CentraleSupélec-CNRS-Université Paris-Sud, Gif-sur-Yvette, France.

Part of this work was supported by the FP7 Network of Excellence in Wireless COMMunications NEWCOM# and by the WWTF under grant ICT12-054 (TINCOIN).

In 2003, Dhillon *et al.* [9] adopted an information-theoretic approach to biclustering. They used mutual information to characterize the quality of a biclustering. Specifically, for the special case when the underlying matrix represents the joint probability distribution of two discrete random variables X and Y , i.e., $a_{nm} = P\{X = n, Y = m\}$, their goal was to find functions $f: \{1, \dots, N\} \rightarrow \{1, \dots, K\}$ and $g: \{1, \dots, M\} \rightarrow \{1, \dots, L\}$ that maximize $I(f(X); g(Y))$ for specific K and L . In information-theoretic terms this is a single-letter problem. But by using a stationary, memoryless process and defining achievability in the usual Shannon sense, the approach of Dhillon *et al.* can be addressed via information-theoretic tools.

The aim of the present paper is to characterize the achievable region of this *information-theoretic biclustering problem*. Furthermore we provide an extension to more than two stationary, memoryless sources. This distributed information-theoretic biclustering problem offers a formidable mathematical complexity. It is fundamentally different from “classical” distributed source coding problems like distributed lossy compression [10, Chapter 12]. Usually, one aims at reducing redundant information, i.e., information that is transmitted by multiple encoders, as much as possible, while still guaranteeing correct decoding. But in the biclustering problem, we are interested in maximizing precisely this redundant information. In this sense, the biclustering problem is complementary to distributed source coding. Furthermore, it appears to be closely related to hypothesis testing against independence with multiterminal data compression [11], which is not yet solved in general [12], but the case of full side-information is solved [13]. We also point out that the extension to multiple variables contains the Körner-Marton problem [14], which implies that in general Berger-Tung coding is suboptimal.

B. Contributions

We first study the case of two sources in Section II and provide an outer and an inner bound on the achievable region. The outer bound follows from standard information-theoretic manipulations, while the inner bound uses more involved methods developed for solving hypothesis testing problems, specifically [11]. We argue why the standard random coding arguments are not sufficient to show an achievability result for this problem. In Section II-D we extensively study the doubly symmetric binary source as an example. In order to perform this analysis, we require stronger cardinality bounds, than the ones usually obtained using the convex cover method [10, Appendix C]. To achieve this, we combine the convex cover method, the perturbation method and leverage ideas similar to [15], allowing us to deal only with the extreme points of the achievable region. The resulting bounds, proved in Appendix C, allow us to choose binary auxiliaries for binary sources. Based on numerical evidence we then argue that there is a gap between the outer and the inner bound for a doubly symmetric binary source. In Section III we extend both bounds to the case of multiple sources, which additionally requires a binning strategy for the achievability part. In Section III-B we investigate the *CEO problem under a mutual information constraint*, a special case of the information-theoretic biclustering problem with multiple sources. We show that it is equivalent to classical multiterminal lossy source coding under logarithmic loss distortion. By leveraging this equivalence, we obtain tight bounds for a special case using results from [16].

C. Related Work

The *information bottleneck* (IB) method introduced by Tishby *et al.* [17] can be interpreted as the unidirectional variant of information-theoretic biclustering [9]. The corresponding Shannon-theoretic bottleneck problem of maximizing *relevance* $\mu = I(f(X^n); Z^n)/n$ while limiting the rate $\log \|f\| \leq nR$ was investigated in [18]. We show that this information bottleneck problem is equivalent to lossy source coding with *logarithmic loss distortion* [16], which is solved in terms of a single-letter description. The maximum attainable relevance is given by

$$\mu(R) = \max_{\substack{U : I(U; X) \leq R \\ U \leftrightarrow X \leftrightarrow Z}} I(U; Z). \quad (1)$$

Interestingly, the function (1) is the solution to a variety of different problems in information theory. As mentioned in [18], (1) is the solution to the problem of loss-less source coding with one helper [19], [20]. Witsenhausen and Wyner [21] investigated a lower bound for a conditional entropy when simultaneously requiring another conditional entropy to fall below a threshold. Their work was a generalization of an earlier result [22] and furthermore related to [19], [23]–[25]. The conditional entropy bound in [21] turns out to be an equivalent characterization of (1). Also

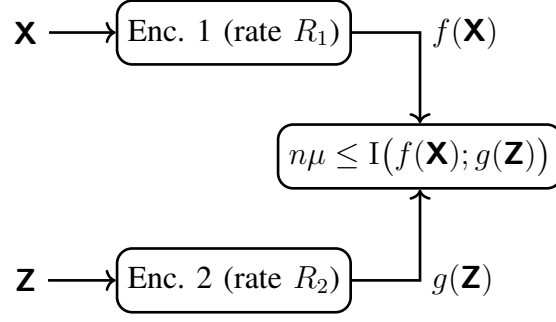


Fig. 1: Biclustering of two memoryless sources.

in the context of gambling in the horse race market, (1) occurs as the maximum incremental growth in wealth when rate-limited side-information is available to the gambler [26, Theorem 3].

Recently, Kumar and Courtade introduced a conjecture [27], [28] concerning Boolean functions that maximize mutual information. Their work was inspired by a similar problem in computational biology [29]. A weaker form of their conjecture [28, Section IV, 2)] corresponds to a zero-rate variant of the binary example studied in Section II-D.

D. Notation and Conventions

We use $\overline{\mathcal{A}}$ to denote the topological closure of a set \mathcal{A} and \mathcal{A}^c for the complement of a set (or event). The symbol $|\mathcal{A}|$ is used for the number of elements in a finite set \mathcal{A} . When there is no possibility of confusion we identify a set with one element with that element, e.g., $\{1, 2, 3\} \setminus 1 = \{2, 3\}$. Let \mathbb{R}_+ be the set of non-negative reals and \mathbb{R}_- the set of non-positive reals. We denote random quantities and their realizations by capital, sans-serif and lowercase letters, respectively. Furthermore, vectors are indicated by bold-face type and have length n , if not otherwise specified. We use subscript and superscript to denote slices of vectors in the usual way, i.e., $\mathbf{x}_l^k \triangleq (\mathbf{x}_i)_{i \in \{l, l+1, \dots, k\}}$ and $\mathbf{x}^k \triangleq \mathbf{x}_1^k$, where x_i is the i -th component of \mathbf{x} . For convenience, we define the sets $\mathcal{K} \triangleq \{1, 2, \dots, K\}$, $\mathcal{J} \triangleq \{1, 2, \dots, J\}$, and $\mathcal{L} \triangleq \{1, 2, \dots, L\}$ for $K, J, L \in \mathbb{N}$. Furthermore let Ω denote the set of all pairs $(\mathcal{A}, \mathcal{B})$, where $\mathcal{A}, \mathcal{B} \subset \mathcal{K}$ are nonempty and disjoint. Note that Ω contains $3^K - 2^{K+1} + 1$ different pairs $(\mathcal{A}, \mathcal{B})$. We also define $\tilde{\Omega}$ as the set of all pairs $(\mathcal{A}, \mathcal{B})$, where $\mathcal{A} \subseteq \mathcal{J}$ and $\mathcal{B} \subseteq \mathcal{K}$ are nonempty. Hence, the set $\tilde{\Omega}$ contains $2^{J+L} - 2^J - 2^L + 1$ elements. Additionally we use subscript sets to denote tuples of real values and their components, e.g., $x_{\mathcal{K}} \in \mathbb{R}^{\mathcal{K}}$ or $x_{\Omega} \in \mathbb{R}^{\Omega}$. Naturally, slices of tuples are indexed by subsets, e.g., $x_{\mathcal{A}}$ for a set $\mathcal{A} \subseteq \mathcal{K}$ is a slice of $x_{\mathcal{K}}$. This notation extends naturally to tuples of vectors, where the subscript indices are separated by a comma, e.g., for $\mathbf{x}_{\mathcal{K}} \in \mathbb{R}^{nK}$, we have $\mathbf{x}_{\mathcal{A}, l}^k \in \mathbb{R}^{(k-l+1)|\mathcal{A}|}$. Random variables are assumed to be supported on finite sets. Unless stated otherwise, we use the same letter for the random variable and for its support set, e.g., the random variable X takes values in \mathcal{X} . Given a random variable X , we write $p_X \in \mathcal{P}(\mathcal{X})$ for its probability mass function (pmf), where $\mathcal{P}(\mathcal{X})$ is the set of all pmfs on \mathcal{X} . We use $\tilde{X} \sim p_X$ or $\tilde{X} \simeq X$ to indicate that \tilde{X} and X have the same distribution. We use the notation of [30, Chapter 2] for information-theoretic quantities. All logarithms in this paper are to base e and therefore all information theoretic quantities are measured in nats. The notation $h_0(p) \triangleq -p \log p - (1-p) \log(1-p)$ is used for the binary entropy function and $a * b \triangleq a(1-b) + (1-a)b$ is the binary convolution operation. The symbol \oplus stands for binary addition and $X \sim \mathcal{B}(p)$ is used to denote a Bernoulli distribution with parameter p . The notation $X \ominus Y \ominus Z$ indicates that X , Y , and Z form a Markov chain in this order. When generating codebooks we will assume that the codebook size is an integer to keep the notation simple. Our results are heavily based on the notion of robust typicality [31], also used in [11]. For convenience, the necessary notation and relevant results on types and typicality are summarized in Appendix A.

II. BICLUSTERING WITH TWO SOURCES

A. Problem Statement

In this section we will introduce the *information-theoretic biclustering problem* (or *biclustering problem* for short) with two sources and provide bounds on its achievable region. A schematic overview of the problem is presented in Figure 1. Let (X, Z) be two random variables. The random vectors (\mathbf{X}, \mathbf{Z}) consist of n i.i.d. copies of (X, Z) .

Given a block length $n \in \mathbb{N}$ and coding rates $R_1, R_2 \in \mathbb{R}$, an (n, R_1, R_2) code (f, g) consists of two functions $f: \mathcal{X}^n \rightarrow \mathcal{M}_1$ and $g: \mathcal{Z}^n \rightarrow \mathcal{M}_2$ such that the finite sets \mathcal{M}_k satisfy $\log |\mathcal{M}_k| \leq nR_k$, $k \in \{1, 2\}$. Thus, the coding rates R_k , $k \in \{1, 2\}$, limit the complexity of the encoders.

Definition 1. For an (n, R_1, R_2) code (f, g) , we define the co-information of f and g as

$$\Theta(f; g) \triangleq \frac{1}{n} I(f(\mathbf{X}); g(\mathbf{Z})). \quad (2)$$

This co-information serves as a measure of the mutual relevance of the two encodings $f(\mathbf{X})$ and $g(\mathbf{Z})$. In contrast to rate-distortion theory, we do not require a specific distortion measure; rather, we quantify the quality of a code in pure information-theoretic terms, namely via mutual information. The idea is to find functions f and g that extract a compressed version of the common randomness in observed data \mathbf{X} and \mathbf{Z} .

Definition 2. A triple $(\mu, R_1, R_2) \in \mathbb{R}^3$ is achievable in the biclustering problem if for some $n \in \mathbb{N}$ there exists an (n, R_1, R_2) code (f, g) such that

$$\Theta(f; g) \geq \mu. \quad (3)$$

The achievable region $\overline{\mathcal{R}}$ in the biclustering problem is defined as the closure of the set \mathcal{R} of achievable triples.

We note that stochastic encodings cannot enlarge the achievable region. Assume that the triple (μ, R_1, R_2) can be achieved using a stochastic encoding. The fact that any stochastic encoding can be represented as a convex combination of deterministic encodings implies that at least one of these deterministic encodings also achieves (μ, R_1, R_2) .

B. Outer Bounds

We first provide outer bounds for the set of achievable triples in biclustering with two sources.

Theorem 3. We have $\mathcal{R} \subseteq \mathcal{R}_o \subseteq \mathcal{R}'_o$, where the two regions \mathcal{R}_o and \mathcal{R}'_o are given by

$$\mathcal{R}_o \triangleq \{(\mu, R_1, R_2) : R_1 \geq I(U; X), R_2 \geq I(V; Z), \mu \leq I(V; Z) + I(U; X) - I(UV; XZ)\}, \quad (4)$$

$$\mathcal{R}'_o \triangleq \{(\mu, R_1, R_2) : R_1 \geq I(U; X), R_2 \geq I(V; Z), \mu \leq \min(I(U; Z), I(V; X))\}, \quad (5)$$

with U and V any pair of random variables satisfying $U \circlearrowleft X \circlearrowleft Z$ and $X \circlearrowleft Z \circlearrowleft V$.

The proof of this result is provided in Appendix B. The regions \mathcal{R}_o and \mathcal{R}'_o are both convex since a time-sharing variable can be incorporated into U and V . Furthermore, \mathcal{R}'_o remains unchanged when U and V are required to satisfy the complete Markov chain $U \circlearrowleft X \circlearrowleft Z \circlearrowleft V$. In fact, \mathcal{R}'_o can be obtained by setting $V = Z$ (respectively $U = X$) and applying the outer bound for the information bottleneck problem (see (1)).

The numerical computation of the outer bounds requires the cardinalities of the auxiliary random variables to be bounded. We therefore complement Theorem 3 with the following result, whose proof is provided in Appendix C.

Proposition 4. We have $\mathcal{R}_o = \text{conv}(\mathcal{S}_o)$ and $\mathcal{R}'_o = \text{conv}(\mathcal{S}'_o)$, where the regions \mathcal{S}_o and \mathcal{S}'_o are defined similarly as \mathcal{R}_o and \mathcal{R}'_o , respectively, but with the additional cardinality bounds $|\mathcal{U}| \leq |\mathcal{X}|$ and $|\mathcal{V}| \leq |\mathcal{Z}|$.

The cardinality bounds in this result are tighter than the usual bounds obtained with the convex cover method [10, Appendix C], where the cardinality has to be increased by one. Thus, when dealing with the binary case in Section II-D binary auxiliaries will be sufficient. The smaller cardinalities come at the price of convexification in Proposition 4 since the regions \mathcal{S}_o and \mathcal{S}'_o are not necessarily convex.

C. Inner Bound

Stating an inner bound, i.e., an achievable region, for the biclustering problem is difficult since the standard tools of typicality coding and the Markov lemma [32] seem insufficient to bound the mutual information between two encodings; this is due to the fact that, contrary to distortion, mutual information involves only the (joint) distribution induced by the encoding functions and not the encoder output values. To prove the following theorem, we thus make heavy use of the theory of types [33] and tools developed in [11] in the context of hypothesis testing problems.

To this end, we establish an equivalence between the the biclustering problem and the characterization of error exponents for hypothesis testing problems; this equivalence allows us to use techniques from [11], specifically [11, Theorem 6] and [11, Lemma 8]. Since these results were stated in [11] without proof and since we use them in a different context, we offer self-contained proofs for the sake of completeness.

Theorem 5. *We have $\mathcal{R}_i \subseteq \overline{\mathcal{R}}$ where*

$$\mathcal{R}_i \triangleq \{(\mu, R_1, R_2) : R_1 \geq I(U; X), R_2 \geq I(V; Z), \mu \leq I(U; V)\}, \quad (6)$$

with auxiliary random variables U, V satisfying $U \circlearrowleft X \circlearrowleft Z \circlearrowleft V$.

The proof of this results is stated in Section II-E. Interestingly, the outer bound \mathcal{R}_o and the inner bound \mathcal{R}_i would coincide if the Markov condition $U \circlearrowleft X \circlearrowleft Z \circlearrowleft V$ were imposed in the definition of \mathcal{R}_o since then $I(V; Z) + I(U; X) - I(UV; XZ) = I(U; V) - I(U; V|XZ) = I(U; V)$.

Employing a binning scheme does not increase the achievable region. The intuition is that binning reduces redundant information transmitted by both encoders, whereas in information-theoretic biclustering this quantity should actually be maximized. A tight bound on the achievable region is obtained using common information [34]. Theorem 3 entails $\mu \leq \min(R_1, R_2)$ for any achievable point (μ, R_1, R_2) . Now let $Y = \zeta_1(X) = \zeta_2(Z)$ be a common part of X and Z . With $U = V = Y$, Theorem 5 implies that $(H(Y), H(Y), H(Y))$ is achievable. Using time-sharing with the trivially achievable point $(0, 0, 0)$ we see that the inner bound is tight if $\mu \leq H(Y)$.

We next improve the inner bound \mathcal{R}_i via convexification. Furthermore, we incorporate the same cardinality bounds as for the outer bound in Proposition 4, thereby enabling us to restrict to binary variables in Section II-D.

Proposition 6. *We have $\mathcal{S}'_i \triangleq \text{conv}(\mathcal{S}_i) = \text{conv}(\mathcal{R}_i) \subseteq \overline{\mathcal{R}}$ where \mathcal{S}_i is defined similarly as \mathcal{R}_i , but with the additional cardinality bounds $|\mathcal{U}| \leq |\mathcal{X}|$, and $|\mathcal{V}| \leq |\mathcal{Z}|$. Furthermore, \mathcal{S}'_i can be explicitly expressed as*

$$\mathcal{S}'_i = \{(\mu, R_1, R_2) : R_1 \geq I(U; X|Q), R_2 \geq I(V; Z|Q), \mu \leq I(U; V|Q)\}, \quad (7)$$

where U, V , and Q are random variables such that $p_{X,Z,U,V,Q} = p_{X,Z} p_{U|X,Q} p_{V|Z,Q} p_Q$, $|\mathcal{U}| \leq |\mathcal{X}|$, $|\mathcal{V}| \leq |\mathcal{Z}|$, and $|\mathcal{Q}| \leq 3$.

The proof of this result is given in Appendix E.

D. Example: Binary Symmetric Sources

Let (X, Z) be a doubly symmetric binary source [10, Example 10.1] with parameter p , i.e., $X \sim \mathcal{B}(\frac{1}{2})$ is a Bernoulli random variable with parameter $\frac{1}{2}$, $N \sim \mathcal{B}(p)$, and $Z \triangleq X \oplus N$. We first show that the inner bound \mathcal{S}'_i and the outer bound \mathcal{R}'_o do not coincide.

Proposition 7. *For the doubly symmetric binary source, $\mathcal{S}'_i \neq \mathcal{R}'_o$.*

Proof: With $U = X \oplus N_1$ and $V = Z \oplus N_2$, where $N_1, N_2 \sim \mathcal{B}(p)$ are independent of (X, Z) and of each other, it follows that $(\mu, R, R) \triangleq (\log 2 - h_0(\alpha * p), \log 2 - h_0(\alpha), \log 2 - h_0(\alpha)) \in \mathcal{R}'_o$ for some $\alpha \in (0, \frac{1}{2})$. Assuming $(\mu, R, R) \in \mathcal{S}'_i$ and choosing U, V , and Q according to Proposition 6 implies that

$$H(X|UQ) \geq h_0(\alpha), \quad (8)$$

$$H(Z|VQ) \geq h_0(\alpha), \quad (9)$$

$$I(U; V|Q) \geq \log 2 - h_0(\alpha * p). \quad (10)$$

Applying Mrs. Gerber's Lemma (MGL) [22, Theorem 1] yields

$$H(X|VQ) \geq h_0(h_b^{-1}(H(Z|VQ)) * p) \geq h_0(\alpha * p), \quad (11)$$

where the second inequality follows from (9). Thus, $I(X; V|Q) \leq \log 2 - h_0(\alpha * p)$ and furthermore $I(X; V|Q) \geq I(U; V|Q)$ due to the Markov chain $U \circlearrowleft (X, Q) \circlearrowleft V$. These two inequalities in combination with (10) imply $I(X; V|Q) = I(U; V|Q)$, or equivalently $I(X; V|UQ) = 0$, which corresponds to the Markov chain property $X \circlearrowleft (U, Q) \circlearrowleft V$. We can therefore write the joint pmf of (U, X, V, Q) in two ways

$$p_{U,X,V,Q}(u, x, v, q) = p_X(x) p_Q(q) p_{U|X,Q}(u|x, q) p_{V|X,Q}(v|x, q) \quad (12)$$

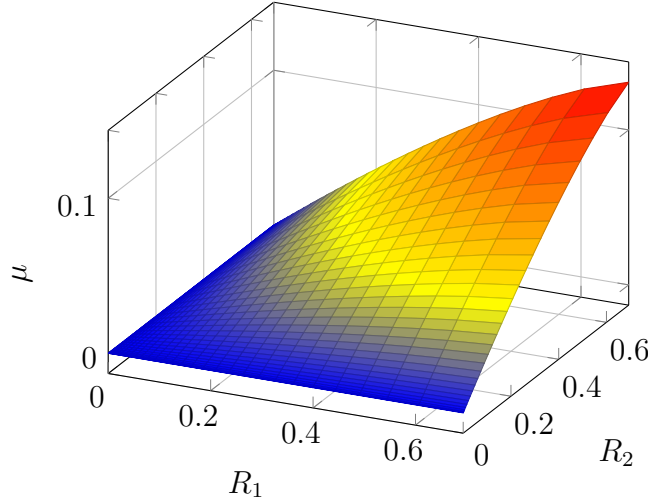


Fig. 2: Boundary of \mathcal{S}_b for $p = \frac{1}{4}$.

$$= p_X(x)p_Q(q)p_{U|X,Q}(u|x,q)p_{V|U,Q}(v|u,q). \quad (13)$$

Assume without loss of generality that $p_Q(q) > 0$ for all $q \in \mathcal{Q}$.

If $p_{U|X,Q}(u|x,q) > 0$ for $(u,x,q) \in \{0,1\} \times \{0,1\} \times \mathcal{Q}$, then (13) necessitates

$$p_{V|U,Q}(v|u,q) = p_{V|X,Q}(v|x,q) \quad (14)$$

for $v \in \{0,1\}$. Next, we partition \mathcal{Q} into the three disjoint subsets

$$\mathcal{Q}_1 \triangleq \{q \in \mathcal{Q} : U = X|Q = q \text{ or } U = 1 \oplus X|Q = q\} \quad (15)$$

$$\mathcal{Q}_2 \triangleq \{q \in \mathcal{Q} : U = 0|Q = q \text{ or } U = 1|Q = q\} \quad (16)$$

$$\mathcal{Q}_3 \triangleq \{q \in \mathcal{Q} : p_{U|X,Q}(0|x,q) > 0 \text{ and } p_{U|X,Q}(1|x,q) > 0 \text{ for some } x \in \{0,1\}\}. \quad (17)$$

Given $q \in \mathcal{Q}_3$, we apply (14) twice and obtain

$$p_{V|U,Q}(v|0,q) = p_{V|X,Q}(v|x,q) = p_{V|U,Q}(v|1,q), \quad (18)$$

i.e., $I(U;V|Q = q) = 0$, which is also true for $q \in \mathcal{Q}_2$. We can then develop (10) as

$$\log 2 - h_0(\alpha * p) \leq I(U;V|Q) = P\{Q \in \mathcal{Q}_1\} I(X;Z) = P\{Q \in \mathcal{Q}_1\} (\log 2 - h_0(p)). \quad (19)$$

On the other hand we obtain from (8) that

$$h_0(\alpha) \leq H(X|UQ) \leq P\{Q \in (\mathcal{Q}_2 \cup \mathcal{Q}_3)\} \log 2 = (1 - P\{Q \in \mathcal{Q}_1\}) \log 2. \quad (20)$$

Combination of the previous two inequalities leads to

$$\frac{\log 2 - h_0(\alpha * p)}{\log 2 - h_0(p)} \leq P\{Q \in \mathcal{Q}_1\} \leq 1 - \frac{h_0(\alpha)}{\log 2}, \quad (21)$$

which is a contradiction since $\frac{\log 2 - h_0(\alpha * p)}{\log 2 - h_0(p)} > 1 - \frac{h_0(\alpha)}{\log 2}$. ■

Let the region \mathcal{S}_b be defined as

$$\mathcal{S}_b \triangleq \bigcup_{0 \leq \alpha, \beta \leq \frac{1}{2}} \{(\mu, R_1, R_2) : R_1 \geq \log 2 - h_0(\alpha), R_2 \geq \log 2 - h_0(\beta), \mu \leq \log 2 - h_0(\alpha * p * \beta)\}.$$

By choosing $U = X \oplus N_1$ and $V = Z \oplus N_2$, where $N_1 \sim \mathcal{B}(\alpha)$ and $N_2 \sim \mathcal{B}(\beta)$ are independent of (X, Z) and of each other, it follows that $\mathcal{S}_b \subseteq \mathcal{S}_i$. To illustrate the tradeoff between complexity (R_1, R_2) and relevance (μ) , the boundary of \mathcal{S}_b is depicted for $p = \frac{1}{4}$ in Figure 2.

Based on numerical experiments, we formulate the following conjecture.

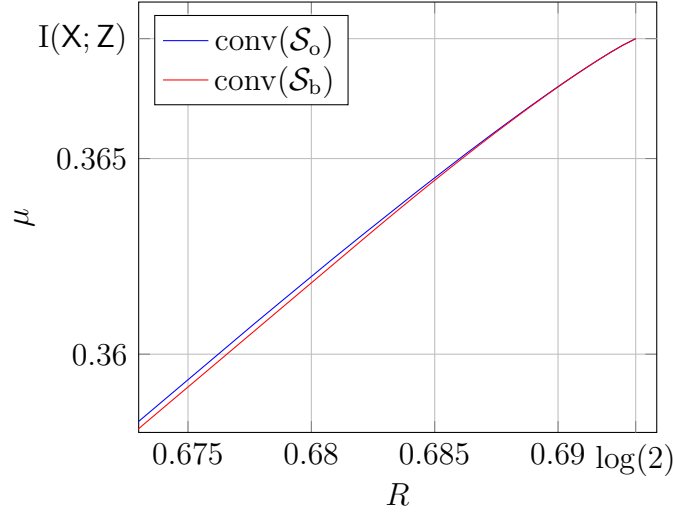


Fig. 3: Outer boundaries of bounds on \bar{R} for $p = 0.1$

Conjecture 8. *Given two binary variables U and V satisfying $U \circlearrowleft X \circlearrowleft Z \circlearrowleft V$, there exist parameters $0 \leq \alpha, \beta \leq \frac{1}{2}$ such that*

$$I(X; U) \geq \log 2 - h_0(\alpha) \quad (22)$$

$$I(Z; V) \geq \log 2 - h_0(\beta) \quad (23)$$

$$I(U; V) \leq \log 2 - h_0(\alpha * p * \beta). \quad (24)$$

Based on Conjecture 8, we can provide numerical evidence that $\mathcal{S}'_i \neq \mathcal{R}_o$. If Conjecture 8 were true, then $\mathcal{S}_b = \mathcal{S}_i$ and hence $\text{conv}(\mathcal{S}_b) = \text{conv}(\mathcal{S}_i) = \mathcal{S}'_i$. We will restrict our discussion to the case $R_1 = R_2 = R$, which leads to a two-dimensional subset of the achievable region. The corresponding subset of $\text{conv}(\mathcal{S}_b)$ can be numerically evaluated. By randomly sampling the binary probability mass functions that satisfy the Markov constraints in Theorem 3, we obtained a strictly larger outer bound.

Based on the cardinality bounds in Propositions 4 and 6 we restrict the auxiliaries U and V to be binary. Using Conjecture 8, we observe that a binary symmetric channel also suffices for \mathcal{S}_i . We thus obtain the parameterized outer boundary of \mathcal{S}_i by evaluating for each (symmetric) rate $\tilde{R}(\alpha) = \log(2) - h_0(\alpha)$ the respective relevance bound

$$\tilde{\mu}_i(\alpha) = \log(2) - h_0(\alpha * p * \alpha). \quad (25)$$

To obtain a bound on $\text{conv}(\mathcal{S}_i)$ we numerically compute the upper concave envelope of $\tilde{\mu}_i(\tilde{R})$. In order to optimize \mathcal{S}_o , we randomly sampled the binary probability mass functions that satisfy the Markov constraints in Theorem 3 (but not necessarily the long Markov chain $U \circlearrowleft X \circlearrowleft Z \circlearrowleft V$). Again, upper concave envelope was computed numerically to obtain the outer boundary of $\text{conv}(\mathcal{S}_o)$.

Figure 3 shows the resulting outer boundaries for $p = 0.1$ in the vicinity of $R = \log(2)$. Albeit small, there is clearly a gap between $\text{conv}(\mathcal{S}_b)$ and $\text{conv}(\mathcal{S}_o)$, showing that the bounds are not tight. We firmly believe that the a tight characterization of the achievable region requires an improved outer bound. However, it appears very difficult to find a manageable outer bound based on the full Markov chain $U \circlearrowleft X \circlearrowleft Z \circlearrowleft V$.

E. Proof of Theorem 5

To prove Theorem 5, we first provide a slightly adapted version of [11, Lemma 8] (see Appendix D for the proof).

Lemma 9. *For the Markov chain $U \circlearrowleft X \circlearrowleft Z \circlearrowleft V$ and any $\varepsilon > 0$, we can choose sufficiently large $n, M_1, M_2 \in \mathbb{N}$ satisfying*

$$\exp(nI(U; X)) < M_1 \leq \exp(n(I(U; X) + \varepsilon)), \quad (26)$$

$$\exp(nI(V; Z)) < M_2 \leq \exp(n(I(V; Z) + \varepsilon)), \quad (27)$$

such that for sufficiently small $\delta > 0$ we have

$$\sum_{i=1}^{M_1} \sum_{j=1}^{M_2} \mathbb{1}_{\mathcal{T}_{[\mathbf{UV}] \delta}^n(\mathbf{u}_i, \mathbf{v}_j)} \cdot \mathbb{P}\{(\mathbf{X}, \mathbf{Z}) \in \mathcal{C}_i \times \mathcal{D}_j\} \geq 1 - \varepsilon \quad (28)$$

and

$$\sum_{i=1}^{M_1} \sum_{j=1}^{M_2} \mathbb{1}_{\mathcal{T}_{[\mathbf{UV}] \delta}^n(\mathbf{u}_i, \mathbf{v}_j)} \leq \exp(n(\mathbf{I}(\mathbf{UV}; \mathbf{XZ}) + \varepsilon)), \quad (29)$$

for some $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_{M_1} \in \mathcal{T}_{[\mathbf{U}] \delta}^n$, $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{M_2} \in \mathcal{T}_{[\mathbf{V}] \delta}^n$, mutually disjoint sets $\mathcal{C}_i \subseteq \mathcal{T}_{[\mathbf{X}|\mathbf{U}] \delta}^n(\mathbf{u}_i)$, $i = 1, \dots, M_1$, and $\mathcal{D}_j \subseteq \mathcal{T}_{[\mathbf{Z}|\mathbf{V}] \delta}^n(\mathbf{v}_j)$, $j = 1, \dots, M_2$.

Furthermore, we will need the following set of random variables.

Definition 10. For random variables $(\mathbf{U}, \mathbf{X}, \mathbf{Z}, \mathbf{V})$, and $\delta \geq 0$ define the set of random variables

$$\mathcal{L}_\delta(\mathbf{U}, \mathbf{X}, \mathbf{Z}, \mathbf{V}) \triangleq \{\tilde{\mathbf{U}}, \tilde{\mathbf{X}}, \tilde{\mathbf{Z}}, \tilde{\mathbf{V}} : (\tilde{\mathbf{U}}, \tilde{\mathbf{X}}) \in \mathcal{T}_{[\mathbf{UX}] \delta}, (\tilde{\mathbf{V}}, \tilde{\mathbf{Z}}) \in \mathcal{T}_{[\mathbf{VZ}] \delta}, (\tilde{\mathbf{U}}, \tilde{\mathbf{V}}) \in \mathcal{T}_{[\mathbf{UV}] \delta}\} \quad (30)$$

and let $\mathcal{L}(\mathbf{U}, \mathbf{X}, \mathbf{Z}, \mathbf{V}) \triangleq \mathcal{L}_0(\mathbf{U}, \mathbf{X}, \mathbf{Z}, \mathbf{V})$.

Note that $\mathcal{L}_\delta(\mathbf{U}, \mathbf{X}, \mathbf{Z}, \mathbf{V}) \subseteq \mathcal{L}_{\delta'}(\mathbf{U}, \mathbf{X}, \mathbf{Z}, \mathbf{V})$ for $\delta \leq \delta'$.

Select \mathbf{U} and \mathbf{V} according to (6). Fix $\varepsilon > 0$, let M_1 and M_2 satisfy (26) and (27) and let $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_{M_1} \in \mathcal{T}_{[\mathbf{U}] \delta}^n$, $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_{M_1}$, $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{M_2} \in \mathcal{T}_{[\mathbf{V}] \delta}^n$, and $\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_{M_2}$ be as given in Lemma 9 for a sufficiently small $\delta > 0$ and n large enough. Define the sets $\mathcal{C} \triangleq \bigcup_{i=1}^{M_1} \mathcal{C}_i$ and $\mathcal{D} \triangleq \bigcup_{j=1}^{M_2} \mathcal{D}_j$ and consider the code

$$f(\mathbf{x}) \triangleq \begin{cases} i, & \mathbf{x} \in \mathcal{C}_i, \\ 0, & \mathbf{x} \notin \mathcal{C}, \end{cases} \quad g(\mathbf{z}) \triangleq \begin{cases} j, & \mathbf{z} \in \mathcal{D}_j, \\ 0, & \mathbf{z} \notin \mathcal{D}. \end{cases} \quad (31)$$

For large enough n , this is an $(n, \mathbf{I}(\mathbf{U}; \mathbf{X}) + 2\varepsilon, \mathbf{I}(\mathbf{V}; \mathbf{Z}) + 2\varepsilon)$ code. We now need to analyze $\Theta(f; g)$. Let $\bar{\mathbf{X}}$ and $\bar{\mathbf{Z}}$ be random variables distributed according to $p_{\bar{\mathbf{X}}, \bar{\mathbf{Z}}}(x, z) = p_{\mathbf{X}}(x)p_{\mathbf{Z}}(z)$, i.e., with the same marginals as \mathbf{X} and \mathbf{Z} , but independent. Defining $\mathbf{W}_1 \triangleq f(\mathbf{X})$, $\mathbf{W}_2 \triangleq g(\mathbf{Z})$, $\bar{\mathbf{W}}_1 \triangleq f(\bar{\mathbf{X}})$, $\bar{\mathbf{W}}_2 \triangleq g(\bar{\mathbf{Z}})$, and $\mathcal{F} \triangleq \{(i, j) : (\mathbf{u}_i, \mathbf{v}_j) \in \mathcal{T}_{[\mathbf{UV}] \delta}^n\}$, we have¹

$$n \Theta(f; g) = \mathbf{I}(f(\mathbf{X}); g(\mathbf{Z})) = \sum_{i,j} \mathbb{P}\{\mathbf{W}_1 = i, \mathbf{W}_2 = j\} \log \frac{\mathbb{P}\{\mathbf{W}_1 = i, \mathbf{W}_2 = j\}}{\mathbb{P}\{\mathbf{W}_1 = i\} \mathbb{P}\{\mathbf{W}_2 = j\}} \quad (32)$$

$$\begin{aligned} &= \sum_{i,j \in \mathcal{F}} \mathbb{P}\{\mathbf{W}_1 = i, \mathbf{W}_2 = j\} \log \frac{\mathbb{P}\{\mathbf{W}_1 = i, \mathbf{W}_2 = j\}}{\mathbb{P}\{\mathbf{W}_1 = i\} \mathbb{P}\{\mathbf{W}_2 = j\}} \\ &\quad + \sum_{i,j \in \mathcal{F}^c} \mathbb{P}\{\mathbf{W}_1 = i, \mathbf{W}_2 = j\} \log \frac{\mathbb{P}\{\mathbf{W}_1 = i, \mathbf{W}_2 = j\}}{\mathbb{P}\{\mathbf{W}_1 = i\} \mathbb{P}\{\mathbf{W}_2 = j\}} \end{aligned} \quad (33)$$

$$\geq p_{\mathcal{F}} \log \frac{p_{\mathcal{F}}}{\bar{p}_{\mathcal{F}}} + (1 - p_{\mathcal{F}}) \log \frac{1 - p_{\mathcal{F}}}{1 - \bar{p}_{\mathcal{F}}} \quad (34)$$

where the last step follows from the log-sum inequality [30, Theorem 2.7.1] with the shorthand notation $p_{\mathcal{F}} \triangleq \mathbb{P}\{(\mathbf{W}_1, \mathbf{W}_2) \in \mathcal{F}\}$ and $\bar{p}_{\mathcal{F}} \triangleq \mathbb{P}\{(\bar{\mathbf{W}}_1, \bar{\mathbf{W}}_2) \in \mathcal{F}\}$. The lower bound (34) can be developed further as

$$p_{\mathcal{F}} \log \frac{p_{\mathcal{F}}}{\bar{p}_{\mathcal{F}}} + (1 - p_{\mathcal{F}}) \log \frac{1 - p_{\mathcal{F}}}{1 - \bar{p}_{\mathcal{F}}} = -h_0(p_{\mathcal{F}}) - p_{\mathcal{F}} \log \bar{p}_{\mathcal{F}} - (1 - p_{\mathcal{F}}) \log(1 - \bar{p}_{\mathcal{F}}) \quad (35)$$

$$\geq -h_0(p_{\mathcal{F}}) - p_{\mathcal{F}} \log \bar{p}_{\mathcal{F}} \geq -h_0(p_{\mathcal{F}}) - (1 - \varepsilon) \log \bar{p}_{\mathcal{F}} \quad (36)$$

$$\geq -\log(2) - (1 - \varepsilon) \log \bar{p}_{\mathcal{F}}, \quad (37)$$

¹Dealing with the quantity $\mathbf{I}(f(\mathbf{X}); g(\mathbf{Z}))$ using the techniques from [11] in addition to the theory of types seems necessary as typicality alone is insufficient. For instance, one could simply choose $f \equiv i$ and $g \equiv j$ constant with $(i, j) \in \mathcal{F}$. In this case, joint typicality holds with probability one but $\Theta(f; g) = 0$.

where we have used (28). For each $i \in \mathcal{M}_1$ and $j \in \mathcal{M}_2$ define

$$\mathcal{S}(i, j) \triangleq \{\mathbf{u}_i\} \times \mathcal{C}_i \times \mathcal{D}_j \times \{\mathbf{v}_j\} \quad (38)$$

and

$$\mathbb{S} \triangleq \bigcup_{i,j \in \mathcal{F}} \mathcal{S}(i, j). \quad (39)$$

Pick any $(\hat{\mathbf{u}}, \hat{\mathbf{x}}, \hat{\mathbf{z}}, \hat{\mathbf{v}}) \in \mathbb{S}$ and let $\hat{\mathbf{U}}, \hat{\mathbf{X}}, \hat{\mathbf{Z}},$ and $\hat{\mathbf{V}}$ be the type variables corresponding to $\hat{\mathbf{u}}, \hat{\mathbf{x}}, \hat{\mathbf{z}},$ and $\hat{\mathbf{v}}$ (Definition 21). We then have

$$p_{\bar{\mathbf{X}}, \bar{\mathbf{Z}}}(\hat{\mathbf{x}}, \hat{\mathbf{z}}) = \exp \left(-n(H(\hat{\mathbf{X}}\hat{\mathbf{Z}}) + D_{\text{KL}}(\hat{\mathbf{X}}\hat{\mathbf{Z}} \parallel \bar{\mathbf{X}}\bar{\mathbf{Z}})) \right) \quad (40)$$

from part 1 of Lemma 23. Let $\kappa(i, j; \hat{\mathbf{U}}, \hat{\mathbf{X}}, \hat{\mathbf{Z}}, \hat{\mathbf{V}})$ be the number of elements in $\mathcal{S}(i, j)$ with type $(\hat{\mathbf{U}}, \hat{\mathbf{X}}, \hat{\mathbf{Z}}, \hat{\mathbf{V}})$, then by part 2 of Lemma 23

$$\kappa(i, j; \hat{\mathbf{U}}, \hat{\mathbf{X}}, \hat{\mathbf{Z}}, \hat{\mathbf{V}}) \leq \exp \left(nH(\hat{\mathbf{X}}\hat{\mathbf{Z}} | \hat{\mathbf{U}}\hat{\mathbf{V}}) \right). \quad (41)$$

Let $\kappa(\hat{\mathbf{U}}, \hat{\mathbf{X}}, \hat{\mathbf{Z}}, \hat{\mathbf{V}})$ be the number of elements of \mathbb{S} with type $(\hat{\mathbf{U}}, \hat{\mathbf{X}}, \hat{\mathbf{Z}}, \hat{\mathbf{V}})$. Then

$$\kappa(\hat{\mathbf{U}}, \hat{\mathbf{X}}, \hat{\mathbf{Z}}, \hat{\mathbf{V}}) = \sum_{(i,j) \in \mathcal{F}} \kappa(i, j; \hat{\mathbf{U}}, \hat{\mathbf{X}}, \hat{\mathbf{Z}}, \hat{\mathbf{V}}) \quad (42)$$

$$\leq \sum_{(i,j) \in \mathcal{F}} \exp \left(nH(\hat{\mathbf{X}}\hat{\mathbf{Z}} | \hat{\mathbf{U}}\hat{\mathbf{V}}) \right) \quad (43)$$

$$\leq \exp \left(n(I(\mathbf{UV}; \mathbf{XZ}) + H(\hat{\mathbf{X}}\hat{\mathbf{Z}} | \hat{\mathbf{U}}\hat{\mathbf{V}}) + \varepsilon) \right), \quad (44)$$

where the last inequality is a consequence of (29). Thus,

$$\bar{p}_{\mathcal{F}} = \sum_{\hat{\mathbf{U}}, \hat{\mathbf{X}}, \hat{\mathbf{Z}}, \hat{\mathbf{V}}} \kappa(\hat{\mathbf{U}}, \hat{\mathbf{X}}, \hat{\mathbf{Z}}, \hat{\mathbf{V}}) \exp \left(-n(H(\hat{\mathbf{X}}\hat{\mathbf{Z}}) + D_{\text{KL}}(\hat{\mathbf{X}}\hat{\mathbf{Z}} \parallel \bar{\mathbf{X}}\bar{\mathbf{Z}})) \right) \quad (45)$$

$$\leq \sum_{\hat{\mathbf{U}}, \hat{\mathbf{X}}, \hat{\mathbf{Z}}, \hat{\mathbf{V}}} \exp \left(-n(k(\hat{\mathbf{U}}, \hat{\mathbf{X}}, \hat{\mathbf{Z}}, \hat{\mathbf{V}}) - \varepsilon) \right) \quad (46)$$

where the sum is over all types that occur in \mathbb{S} and we defined

$$k(\hat{\mathbf{U}}, \hat{\mathbf{X}}, \hat{\mathbf{Z}}, \hat{\mathbf{V}}) \triangleq I(\hat{\mathbf{U}}\hat{\mathbf{V}}; \hat{\mathbf{X}}\hat{\mathbf{Z}}) - I(\mathbf{UV}; \mathbf{XZ}) + D_{\text{KL}}(\hat{\mathbf{X}}\hat{\mathbf{Z}} \parallel \bar{\mathbf{X}}\bar{\mathbf{Z}}). \quad (47)$$

Using a type counting argument (Lemma 22) we can further bound

$$\bar{p}_{\mathcal{F}} \leq (n+1)^{|\mathcal{U}||\mathcal{X}||\mathcal{Z}||\mathcal{V}|} \max_{\hat{\mathbf{U}}, \hat{\mathbf{X}}, \hat{\mathbf{Z}}, \hat{\mathbf{V}}} \exp \left(-n(k(\hat{\mathbf{U}}, \hat{\mathbf{X}}, \hat{\mathbf{Z}}, \hat{\mathbf{V}}) - \varepsilon) \right) \quad (48)$$

where the maximum is over all types occurring in \mathbb{S} . For any type $(\hat{\mathbf{U}}, \hat{\mathbf{X}}, \hat{\mathbf{Z}}, \hat{\mathbf{V}})$ in \mathbb{S} , we have by construction $(\hat{\mathbf{U}}, \hat{\mathbf{X}}, \hat{\mathbf{Z}}, \hat{\mathbf{V}}) \in \mathcal{L}_{\delta}(\mathbf{U}, \mathbf{X}, \mathbf{Z}, \mathbf{V})$. From (48) we can thus conclude

$$\bar{p}_{\mathcal{F}} \leq (n+1)^{|\mathcal{U}||\mathcal{X}||\mathcal{Z}||\mathcal{V}|} \max_{(\tilde{\mathbf{U}}, \tilde{\mathbf{X}}, \tilde{\mathbf{Z}}, \tilde{\mathbf{V}}) \in \mathcal{L}_{\delta}(\mathbf{U}, \mathbf{X}, \mathbf{Z}, \mathbf{V})} \exp \left(-n(k(\tilde{\mathbf{U}}, \tilde{\mathbf{X}}, \tilde{\mathbf{Z}}, \tilde{\mathbf{V}}) - \varepsilon) \right). \quad (49)$$

Combining (37) and (49) we showed for n large enough

$$\Theta(f; g) \geq -\frac{\log(2)}{n} - \frac{1-\varepsilon}{n} \log \bar{p}_{\mathcal{F}} \quad (50)$$

$$\geq -\varepsilon + (1-\varepsilon) \left(\min_{(\tilde{\mathbf{U}}, \tilde{\mathbf{X}}, \tilde{\mathbf{Z}}, \tilde{\mathbf{V}}) \in \mathcal{L}_{\delta}(\mathbf{U}, \mathbf{X}, \mathbf{Z}, \mathbf{V})} k(\tilde{\mathbf{U}}, \tilde{\mathbf{X}}, \tilde{\mathbf{Z}}, \tilde{\mathbf{V}}) - \varepsilon \right) \quad (51)$$

$$\geq -2\varepsilon + (1-\varepsilon) \min_{(\tilde{\mathbf{U}}, \tilde{\mathbf{X}}, \tilde{\mathbf{Z}}, \tilde{\mathbf{V}}) \in \mathcal{L}_{\delta}(\mathbf{U}, \mathbf{X}, \mathbf{Z}, \mathbf{V})} k(\tilde{\mathbf{U}}, \tilde{\mathbf{X}}, \tilde{\mathbf{Z}}, \tilde{\mathbf{V}}) \quad (52)$$

$$\geq \min_{(\tilde{U}, \tilde{X}, \tilde{Z}, \tilde{V}) \in \mathcal{L}_\delta(\mathcal{U}, \mathcal{X}, \mathcal{Z}, \mathcal{V})} k(\tilde{U}, \tilde{X}, \tilde{Z}, \tilde{V}) - (2 + I(\mathbf{X}; \mathbf{Z}))\varepsilon \quad (53)$$

$$= \min_{(\tilde{U}, \tilde{X}, \tilde{Z}, \tilde{V}) \in \mathcal{L}_\delta(\mathcal{U}, \mathcal{X}, \mathcal{Z}, \mathcal{V})} k(\tilde{U}, \tilde{X}, \tilde{Z}, \tilde{V}) - C\varepsilon \quad (54)$$

for some constant C . As $k(\tilde{U}, \tilde{X}, \tilde{Z}, \tilde{V})$ is continuous as a function of $p_{\tilde{U}, \tilde{X}, \tilde{Z}, \tilde{V}}$ and (54) holds for arbitrarily small δ , we obtain for n large enough

$$\Theta(f; g) \geq \min_{(\tilde{U}, \tilde{X}, \tilde{Z}, \tilde{V}) \in \mathcal{L}(\mathcal{U}, \mathcal{X}, \mathcal{Z}, \mathcal{V})} k(\tilde{U}, \tilde{X}, \tilde{Z}, \tilde{V}) - C'\varepsilon \quad (55)$$

for some (larger) constant C' by letting $\delta \rightarrow 0$. For $(\tilde{U}, \tilde{X}, \tilde{Z}, \tilde{V}) \in \mathcal{L}(\mathcal{U}, \mathcal{X}, \mathcal{Z}, \mathcal{V})$, observe that

$$k(\tilde{U}, \tilde{X}, \tilde{Z}, \tilde{V}) = I(\tilde{U}\tilde{V}; \tilde{X}\tilde{Z}) - I(UV; \mathbf{X}\mathbf{Z}) + D_{\text{KL}}(\tilde{X}\tilde{Z} \| \overline{X}\overline{Z}) \quad (56)$$

$$= H(\tilde{U}\tilde{V}) - H(\tilde{U}\tilde{V} | \tilde{X}\tilde{Z}) - H(UV) + H(UV | \mathbf{X}\mathbf{Z}) + D_{\text{KL}}(\tilde{X}\tilde{Z} \| \overline{X}\overline{Z}) \quad (57)$$

$$\stackrel{(a)}{=} H(\tilde{U} | \tilde{X}) + H(\tilde{V} | \tilde{Z}) - H(\tilde{U}\tilde{V} | \tilde{X}\tilde{Z}) + D_{\text{KL}}(\tilde{X}\tilde{Z} \| \overline{X}\overline{Z}) \quad (58)$$

$$= \sum_{u,v} \sum_{x,z} p_{\tilde{U}, \tilde{X}, \tilde{Z}, \tilde{V}}(u, x, z, v) [\log p_{\tilde{U}, \tilde{V}, \tilde{X}, \tilde{Z}}(u, v, x, z) - \log p_{\tilde{U} | \tilde{X}}(u | x) \\ - \log p_{\tilde{V} | \tilde{Z}}(v | z) - \log p_{\overline{X}, \overline{Z}}(x, z)] \quad (59)$$

$$\stackrel{(b)}{=} D_{\text{KL}}(\tilde{U}\tilde{X}\tilde{Z}\tilde{V} \| \overline{U}\overline{X}\overline{Z}\overline{V}) \quad (60)$$

$$= I(\tilde{X}\tilde{U}; \tilde{Z}\tilde{V}), \quad (61)$$

where (a) follows from $(\tilde{U}, \tilde{X}, \tilde{Z}, \tilde{V}) \in \mathcal{L}(\mathcal{U}, \mathcal{X}, \mathcal{Z}, \mathcal{V})$ and $\mathbf{U} \ominus \mathbf{X} \ominus \mathbf{Z} \ominus \mathbf{V}$ and (b) is obtained with random variables \overline{U} and \overline{V} uniquely determined by the requirements $\overline{U} \ominus \overline{X} \ominus \overline{Z} \ominus \overline{V}$, $(\overline{X}, \overline{U}) \simeq (\mathbf{X}, \mathbf{U})$, and $(\overline{Z}, \overline{V}) \simeq (\mathbf{Z}, \mathbf{V})$. Clearly we have

$$\min_{\tilde{U}, \tilde{X}, \tilde{Z}, \tilde{V} \in \mathcal{L}(\mathcal{U}, \mathcal{X}, \mathcal{Z}, \mathcal{V})} I(\tilde{X}\tilde{U}; \tilde{Z}\tilde{V}) \geq I(\mathbf{U}; \mathbf{V}) \quad (62)$$

as $I(\tilde{X}\tilde{U}; \tilde{Z}\tilde{V}) \geq I(\tilde{U}; \tilde{V})$ and for $(\tilde{U}, \tilde{X}, \tilde{Z}, \tilde{V}) \in \mathcal{L}(\mathcal{U}, \mathcal{X}, \mathcal{Z}, \mathcal{V})$ we have $(\tilde{U}, \tilde{V}) \simeq (\mathbf{U}, \mathbf{V})$. Combining (55), (61) and (62) shows that $(\mu, R_1, R_2) \in \mathcal{R}$ with $\mu = I(\mathbf{U}; \mathbf{V}) - C'\varepsilon$, $R_1 = I(\mathbf{X}; \mathbf{U}) + 2\varepsilon$, and $R_2 = I(\mathbf{Z}; \mathbf{V}) + 2\varepsilon$. This completes the proof as ε was arbitrary.

III. BICLUSTERING WITH MULTIPLE SOURCES

A. Problem Statement and Results

In this section we extend the information-theoretic biclustering problem introduced in Section II-A to the case of multiple sources and we provide bounds on the associated achievable region. A schematic illustration of the problem is shown in Figure 4.

Let $\mathbf{X}_{\mathcal{K}}$ be K random variables, taking values in the finite sets $\mathcal{X}_{\mathcal{K}}$. The random vectors $\mathbf{X}_{\mathcal{K}}$ consist of i.i.d. copies of $\mathbf{X}_{\mathcal{K}}$. For $n \in \mathbb{N}$ and $R_{\mathcal{K}} \in \mathbb{R}^K$, an $(n, R_{\mathcal{K}})$ code $f_{\mathcal{K}}$ consists of K functions $f_k: \mathcal{X}_{\mathcal{K}}^n \rightarrow \mathcal{M}_k$, where \mathcal{M}_k is an arbitrary finite set with $\log |\mathcal{M}_k| \leq nR_k$ for each $k \in \mathcal{K}$. The symbol μ_{Ω} refers to a tuple $\mu_{\Omega} \in \mathbb{R}^{\Omega}$, where Ω is the set of all pairs $(\mathcal{A}, \mathcal{B})$ with $\mathcal{A}, \mathcal{B} \subset \mathcal{K}$ nonempty and disjoint.

Definition 11. Consider an $(n, R_{\mathcal{K}})$ code $f_{\mathcal{K}}$ with $U_k \triangleq f_k(\mathbf{X}_{\mathcal{K}})$; for any $(\mathcal{A}, \mathcal{B}) \in \Omega$ we define the co-information of $f_{\mathcal{A}}$ and $f_{\mathcal{B}}$ as

$$\Theta(f_{\mathcal{A}}; f_{\mathcal{B}}) \triangleq \frac{1}{n} I(U_{\mathcal{A}}; U_{\mathcal{B}}). \quad (63)$$

Definition 12. A point $(\mu_{\Omega}, R_{\mathcal{K}})$ is achievable if for some $n \in \mathbb{N}$ there exists an $(n, R_{\mathcal{K}})$ code $f_{\mathcal{K}}$ such that for any $(\mathcal{A}, \mathcal{B}) \in \Omega$

$$\Theta(f_{\mathcal{A}}; f_{\mathcal{B}}) \geq \mu_{\mathcal{A}, \mathcal{B}}. \quad (64)$$

The set of all achievable points is denoted \mathcal{R} and we refer to its closure $\overline{\mathcal{R}}$ as achievable region.

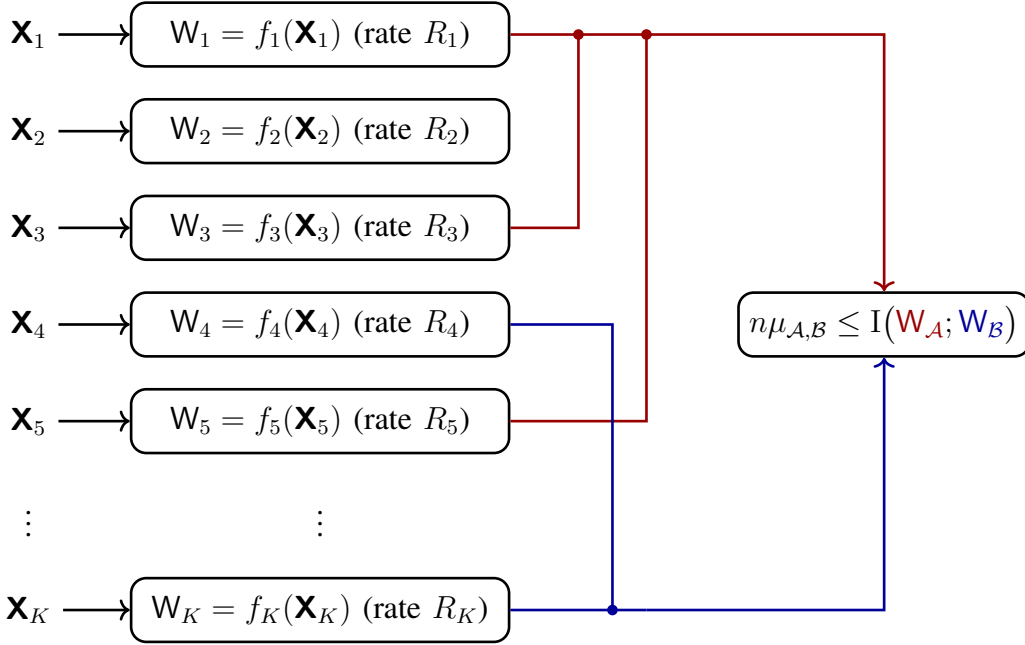


Fig. 4: Biclustering of multiple sources

We first state an outer bound for the achievable region whose proof is provided in Appendix F.

Theorem 13. *We have the outer bounds $\mathcal{R} \subseteq \mathcal{R}_o \subseteq \mathcal{R}'_o$. Here, the region \mathcal{R}'_o is defined as*

$$\mathcal{R}'_o \triangleq \{(\mu_\Omega, R_K) : R_k \geq I(U_k; X_k), k \in \mathcal{K}, \text{ and } \mu_{A,B} \leq I(U_A; X_B), (A, B) \in \Omega\}$$

for some random variables U_K with $U_A \ominus X_A \ominus X_K$ for any $A \subseteq K$. The region \mathcal{R}_o is defined similarly as \mathcal{R}'_o only that the inequality for the relevance $\mu_{A,B}$ is replaced with

$$\mu_{A,B} \leq I(U_A; X_A) + I(U_B; X_B) - I(U_A U_B; X_A X_B). \quad (65)$$

The next results, whose proof is detailed in Section III-C, provides an achievable region.

Theorem 14. *An inner bound for the achievable region is given by $\mathcal{R}_i \subseteq \overline{\mathcal{R}}$ where the region \mathcal{R}_i consists of all points (μ_Ω, R_K) for which there exist random variables U_K satisfying $U_k \ominus X_k \ominus (X_{K \setminus k}, U_{K \setminus k})$ for all $k \in K$ and for all $(A, B) \in \Omega$ there exist subsets $\mathcal{A}_b \subseteq \mathcal{A}_a \subseteq \mathcal{A}$ and $\mathcal{B}_b \subseteq \mathcal{B}_a \subseteq \mathcal{B}$ such that*

$$\sum_{k \in \mathcal{A}'} R_k \geq I(X_{\mathcal{A}'}; U_{\mathcal{A}'} | U_{\mathcal{A}_a \setminus \mathcal{A}'}) \text{ for all } \mathcal{A}' \subseteq \mathcal{A}_a \text{ with } \mathcal{A}' \cap \mathcal{A}_b \neq \emptyset, \quad (66)$$

$$\sum_{k \in \mathcal{B}'} R_k \geq I(X_{\mathcal{B}'}; U_{\mathcal{B}'} | U_{\mathcal{B}_a \setminus \mathcal{B}'}) \text{ for all } \mathcal{B}' \subseteq \mathcal{B}_a \text{ with } \mathcal{B}' \cap \mathcal{B}_b \neq \emptyset, \quad (67)$$

$$\mu_{A,B} \leq I(U_{\mathcal{A}_b}; U_{\mathcal{B}_b}). \quad (68)$$

In contrast to the case of two sources binning does help for $K > 2$ sources. For illustration, consider the case $K = 3$ and assume we are only interested in maximizing $\Theta(f_{\{1,2\}}; f_3)$. Then any information encoded by both f_1 and f_2 is redundant as it does not increase $I(f_1(X_1), f_2(X_2); f_3(X_3))$. The corresponding rate loss can be reduced by a quantize-and-bin scheme (see [19], [32], [35]).

The proof that \mathcal{R}_i is indeed achievable uses typicality coding and binning. The conditions (66) and (67) ensure that $U_{\mathcal{A}_b}$ and $U_{\mathcal{B}_b}$ can be correctly decoded from the output of the encoders \mathcal{A}_a and \mathcal{B}_a , respectively. By (68), $U_{\mathcal{A}_b}$ and $U_{\mathcal{B}_b}$ suffice to ensure that $\mu_{A,B}$ is achievable. Intuitively, the encoders $\mathcal{A}_a \setminus \mathcal{A}_b$ and $\mathcal{B}_a \setminus \mathcal{B}_b$ act as helpers for decoding $U_{\mathcal{A}_b}$ and $U_{\mathcal{B}_b}$, respectively. The special case $\mathcal{A}_b = \mathcal{A}$, $\mathcal{B}_b = \mathcal{B}$ for every $(A, B) \in \Omega$ corresponds to no binning at all, as (66) and (67) then imply $R_k \geq I(X_k; U_k)$ for all $k \in K$. It may seem counterintuitive that the output of the encoders $\mathcal{A}_a \setminus \mathcal{A}_b$ (and analogously $\mathcal{B}_a \setminus \mathcal{B}_b$) is ignored for typicality decoding. However, successful

decoding presupposes (66), which in general does not hold for $\mathcal{A}_a = \mathcal{A}$.

The inner bound in Theorem 14 cannot be tight in general as it contains the Körner-Marton problem [14] as a special case. For $K = 3$ consider $\mathbf{X}_1 \sim \mathcal{B}(\frac{1}{2})$ and $\mathbf{X}_3 \sim \mathcal{B}(p)$ with $p \in (0, 1)$ and $p \neq \frac{1}{2}$. Then define $\mathbf{X}_2 \triangleq \mathbf{X}_1 \oplus \mathbf{X}_3$. The point (μ_Ω, R_K) where $R_3 = \log(2)$, $R_1 = R_2 = H(\mathbf{X}_3) = h_0(p)$, and $\mu_{\mathcal{A}, \mathcal{B}} = 0$ except for $\mu_{\{1,2\},\{3\}} = H(\mathbf{X}_3) = h_0(p)$ is achievable [14, Theorem 1]. However, the quantize-and-bin scheme cannot achieve this point [14, Proposition 1].

Finally, note that in general \mathcal{R}_i is not convex and thus Theorem 14 can be strengthened to $\text{conv}(\mathcal{R}_i) \subseteq \overline{\mathcal{R}}$. However, characterizing $\text{conv}(\mathcal{R}_i)$ using a time-sharing random variable is tedious due to the freedom of choosing the index sets $\mathcal{A}_a, \mathcal{A}_b, \mathcal{B}_a$, and \mathcal{B}_b for each $(\mathcal{A}, \mathcal{B}) \in \Omega$ in Theorem 14.

The following cardinality bounds show that \mathcal{R}_i is computable (see Appendix G for the proof).

Proposition 15. *The region \mathcal{R}_i remains the same if the cardinality bound $|\mathcal{U}_k| \leq |\mathcal{X}_k| + 4^K$ is imposed for every $k \in \mathcal{K}$.*

B. A Special Case: The CEO Problem

In this section we study a special case of the biclustering problem that corresponds to a variant of the CEO problem [36] in which the usual distortion criterion is replaced with mutual information (MI). This problem turns out to be equivalent to the classical CEO problem with *logarithmic loss (log-loss) distortion* [16]. We will show that our inner bound becomes tight in a special case.

We consider a CEO problem under a mutual information constraint where random variables $\mathbf{X}_{\mathcal{J}}$ are encoded to be maximally informative about another set of random variables $\mathbf{Y}_{\mathcal{L}}$ (remember, $\mathcal{J} = \{1, 2, \dots, J\}$ and $\mathcal{L} = \{1, 2, \dots, L\}$). We use the symbol $\nu_{\tilde{\Omega}}$ to denote a tuple $\nu \in \mathbb{R}^{\tilde{\Omega}}$, where $\tilde{\Omega}$ consists of all pairs $(\mathcal{A}, \mathcal{B})$ of nonempty sets $\mathcal{A} \subseteq \mathcal{J}$ and $\mathcal{B} \subseteq \mathcal{L}$.

Definition 16. *A point $(\nu_{\tilde{\Omega}}, R_{\mathcal{J}})$ is MI-achievable if for some $n \in \mathbb{N}$ there exists an $(n, R_{\mathcal{J}})$ code $f_{\mathcal{J}}$ for $\mathbf{X}_{\mathcal{J}}$ such that for all $(\mathcal{A}, \mathcal{B}) \in \tilde{\Omega}$,*

$$\frac{1}{n} \mathbb{I}(\mathbf{U}_{\mathcal{A}}; \mathbf{Y}_{\mathcal{B}}) \geq \nu_{\mathcal{A}, \mathcal{B}}, \quad (69)$$

where $\mathbf{U}_m \triangleq f_m(\mathbf{X}_m)$. Denote the set of all MI-achievable points by \mathcal{R}_{MI} .

The MI-achievable region \mathcal{R}_{MI} is a special case of \mathcal{R} . Indeed, let $\mathbf{X}_{\mathcal{K}} = (\mathbf{X}_{\mathcal{J}}, \mathbf{Y}_{\mathcal{L}})$ with $J + L = K$ and let $\mathbf{Y}_{\mathcal{L}}$ be encoded without compression, i.e., $R_{J+l} = \log|\mathcal{Y}_l|$ for $l \in \mathcal{L}$. Since we are only interested in the information the encodings of $\mathbf{X}_{\mathcal{J}}$ provide about $\mathbf{Y}_{\mathcal{L}}$ we choose $\mu_{\mathcal{A}, \mathcal{B}} = 0$ unless $\mathcal{A} \subseteq \mathcal{J}$ and $\mathcal{B} \subseteq \mathcal{J} + \mathcal{L}$.

Definition 16 also encompasses the information bottleneck problem introduced in [18]. While the definition of a code in [18, Definition 1] differs slightly from Definitions 11 and 12, the achievable region is the same.

We next argue that the CEO problem introduced in Definition 16 is equivalent to the log-loss distortion approach in [16]. Logarithmic loss distortion is defined on the reconstruction alphabet $\mathcal{P}(\mathcal{Y}_{\mathcal{B}}^n)$, the set of all pmfs on $\mathcal{Y}_{\mathcal{B}}^n$ ($\mathcal{B} \subseteq \mathcal{L}$). For $\mathcal{A} \subseteq \mathcal{J}$ and $\mathcal{B} \subseteq \mathcal{L}$ we consider a decoding function $g_{\mathcal{A}, \mathcal{B}}: \mathcal{M}_{\mathcal{A}} \rightarrow \mathcal{P}(\mathcal{Y}_{\mathcal{B}}^n)$ that produces a probabilistic estimate of $\mathbf{Y}_{\mathcal{B}}$ given the output of the encoders \mathcal{A} . The quality of this probabilistic estimate is measured by log-loss distortion. A minor technical difference is that [16] uses conditional entropy (which should be small) whereas we use mutual information (which should be large). Therefore, we prefer to work with log-loss fidelity $\zeta_{\mathcal{B}}$, $\mathcal{B} \subseteq \mathcal{L}$, which is defined as the negative log-loss distortion $d_{\mathcal{B}}$ on $\mathcal{Y}_{\mathcal{B}}^n$ augmented by entropy, i.e.,

$$\zeta_{\mathcal{B}}: \mathcal{Y}_{\mathcal{B}}^n \times \mathcal{P}(\mathcal{Y}_{\mathcal{B}}^n) \rightarrow \mathbb{R}, \quad (70)$$

$$(\mathbf{y}_{\mathcal{B}}, \mathbf{p}) \mapsto H(\mathbf{Y}_{\mathcal{B}}) - d_{\mathcal{B}}(\mathbf{y}_{\mathcal{B}}, \mathbf{p}), \quad (71)$$

where $d_{\mathcal{B}}(\mathbf{y}_{\mathcal{B}}, \mathbf{p}) \triangleq -\frac{1}{n} \log \mathbf{p}(\mathbf{y}_{\mathcal{B}})$ is the logarithmic loss distortion.

Definition 17. *A point $(\nu_{\tilde{\Omega}}, R_{\mathcal{J}})$ is log-loss achievable if for some $n \in \mathbb{N}$ there exists an $(n, R_{\mathcal{J}})$ -code $f_{\mathcal{J}}$ and decoding functions $g_{\mathcal{A}, \mathcal{B}}: \mathcal{M}_{\mathcal{A}} \rightarrow \mathcal{P}(\mathcal{Y}_{\mathcal{B}}^n)$ such that for all $(\mathcal{A}, \mathcal{B}) \in \tilde{\Omega}$*

$$\mathbb{E}[\zeta_{\mathcal{B}}(\mathbf{Y}_{\mathcal{B}}, g_{\mathcal{A}, \mathcal{B}}(\mathbf{U}_{\mathcal{A}}))] \geq \nu_{\mathcal{A}, \mathcal{B}}, \quad (72)$$

where $\mathbf{U}_m \triangleq f_m(\mathbf{X}_m)$. Let \mathcal{R}_{LL} be the set of all log-loss achievable points.

We note that [16] only considers the case where $\mathcal{A} = \mathcal{J}$ and singleton $\mathcal{B} = \{l\}$. To show the equivalence of \mathcal{R}_{MI} and \mathcal{R}_{LL} , we first state an auxiliary lemma that generalizes [16, Lemma 1] (see Appendix H for the proof).

Lemma 18. *For any decoding function $g_{\mathcal{A},\mathcal{B}}$ and code $f_{\mathcal{J}}$, we have*

$$\mathbb{E}[\zeta_{\mathcal{B}}(\mathbf{Y}_{\mathcal{B}}, g_{\mathcal{A},\mathcal{B}}(\mathbf{U}_{\mathcal{A}}))] \leq \frac{1}{n} \mathbb{I}(\mathbf{Y}_{\mathcal{B}}; \mathbf{U}_{\mathcal{A}}), \quad (73)$$

with equality if and only if $g_{\mathcal{A},\mathcal{B}}(u_{\mathcal{A}}) = \mathbf{p}_{\mathbf{Y}_{\mathcal{B}}|\mathbf{U}_{\mathcal{A}}}(\cdot | u_{\mathcal{A}})$.

According to Lemma 18, the MI performance of an encoder-decoder pair is at least as good as its log-loss performance. On the other hand, the optimum can always be achieved when equality holds in (73). This is the basis for the next result.

Lemma 19. *The MI-achievable and log-loss achievable regions are identical, $\mathcal{R}_{\text{LL}} = \mathcal{R}_{\text{MI}}$.*

Proof: First we show $\mathcal{R}_{\text{MI}} \subseteq \mathcal{R}_{\text{LL}}$. Assume we have $(\nu_{\tilde{\Omega}}, R_{\mathcal{J}}) \in \mathcal{R}_{\text{MI}}$. We can thus obtain an $(n, R_{\mathcal{J}})$ -code $f_{\mathcal{J}}$ such that (69) holds. According to Lemma 18, choosing the decoding functions $g_{\mathcal{A},\mathcal{B}}$, as

$$g_{\mathcal{A},\mathcal{B}}(u_{\mathcal{A}}) = \mathbf{p}_{\mathbf{Y}_{\mathcal{B}}|\mathbf{U}_{\mathcal{A}}}(\cdot | u_{\mathcal{A}}) \quad (74)$$

implies $\mathbb{E}[\zeta_{\mathcal{B}}(\mathbf{Y}_{\mathcal{B}}, g_{\mathcal{A},\mathcal{B}}(\mathbf{U}_{\mathcal{A}}))] \geq \nu_{\mathcal{A},\mathcal{B}}$ and thus $(\nu_{\tilde{\Omega}}, R_{\mathcal{J}}) \in \mathcal{R}_{\text{LL}}$. To show $\mathcal{R}_{\text{LL}} \subseteq \mathcal{R}_{\text{MI}}$, assume that $(\nu_{\tilde{\Omega}}, R_{\mathcal{J}}) \in \mathcal{R}_{\text{LL}}$, i.e., there exist $f_{\mathcal{K}}$ and $g_{\mathcal{A},\mathcal{B}}$ such that (72) holds. Lemma 18 then implies (69) and hence $(\nu_{\tilde{\Omega}}, R_{\mathcal{J}}) \in \mathcal{R}$. ■

As a consequence of Lemma 19, the results in [16] directly apply to the CEO problem with mutual information constraint. For example, the CEO problem with J encoders under logarithmic loss distortion [16, Appendix B] can be obtained in Definition 16 by setting $L = 1$ and $\nu_{\mathcal{A},1} = 0$, whenever $\mathcal{A} \neq \mathcal{J}$. The resulting achievable region obtained from \mathcal{R}_i (Theorem 14) consists of all points $(\nu_{\tilde{\Omega}}, R_{\mathcal{J}})$, for which there exists an index subset $\mathcal{A}_b \subseteq \mathcal{J}$ and random variables $\mathbf{U}_{\mathcal{J}}$ with $\mathbf{U}_{\mathcal{J}} \ominus \mathbf{X}_{\mathcal{J}} \ominus \mathbf{Y}_1$ such that

$$\sum_{k \in \mathcal{A}'} R_k \geq \mathbb{I}(\mathbf{X}_{\mathcal{A}'}; \mathbf{U}_{\mathcal{A}'} | \mathbf{U}_{\mathcal{J} \setminus \mathcal{A}'}), \quad \text{for all } \mathcal{A}' \subseteq \mathcal{J} \text{ and } \mathcal{A}' \cap \mathcal{A}_b \neq \emptyset, \quad (75)$$

$$\nu_{\mathcal{J},1} \leq \mathbb{I}(\mathbf{U}_{\mathcal{A}_b}; \mathbf{Y}_1). \quad (76)$$

Remember that $\nu_{\mathcal{A},1} = 0$ whenever $\mathcal{A} \neq \mathcal{J}$. If the random variables $\mathbf{X}_{\mathcal{J}}$ are mutually independent given \mathbf{Y}_1 , we know by [16, Lemma 5] that the inner bound is tight and moreover we can choose $\mathcal{A}_b = \mathcal{J}$. A similar argument holds for the multiterminal source coding problem under logarithmic loss as introduced in [16, Section II]; for that problem, the inner bound obtained from \mathcal{R}_i is also tight due to the results in [16].

C. Proof of Theorem 14

The following lemma (whose proof is provided in Appendix I) incorporates the binning strategy and is the basis for showing that \mathcal{R}_i is indeed an inner bound for the achievable region.

Lemma 20 (Existence of a code). *Let $\varepsilon > 0$, $\mathbf{U}_k \ominus \mathbf{X}_k \ominus (\mathbf{X}_{\mathcal{K} \setminus k}, \mathbf{U}_{\mathcal{K} \setminus k})$ for all $k \in \mathcal{K}$, and $R_{\mathcal{K}} \in \mathbb{R}_+^K$. Then, for sufficiently small $\delta > 0$ and sufficiently large $n \in \mathbb{N}$ we can obtain an $(n, R_{\mathcal{K}} + \varepsilon)$ code $f_{\mathcal{K}}$ with $\mathbf{W}_k \triangleq f_k(\mathbf{X}_k)$ and decoding functions $g_{\mathcal{A},\mathcal{B}}: \mathcal{M}_{\mathcal{A}} \rightarrow \mathcal{U}_{\mathcal{B}}^n$ for each pair of nonempty index sets $\mathcal{B} \subseteq \mathcal{A} \subseteq \mathcal{K}$ such that the following two properties hold:*

- 1) Let $\mathcal{A}_a, \mathcal{A}_b, \mathcal{B}_a, \mathcal{B}_b \subseteq \mathcal{K}$ be arbitrary nonempty subsets of indices with $\mathcal{A}_b \subseteq \mathcal{A}_a$, $\mathcal{B}_b \subseteq \mathcal{B}_a$, and $\mathcal{A}_a \cap \mathcal{B}_a = \emptyset$. If (66) and (67) hold, then, using

$$\mathbb{P} \left\{ (g_{\mathcal{A}_a, \mathcal{A}_b}(\mathbf{W}_{\mathcal{A}_a}), \mathbf{X}_{\mathcal{A}_a}, \mathbf{X}_{\mathcal{B}_a}, g_{\mathcal{B}_a, \mathcal{B}_b}(\mathbf{W}_{\mathcal{B}_a})) \notin \mathcal{T}_{[\mathbf{U}_{\mathcal{A}_b} \mathbf{X}_{\mathcal{A}_a} \mathbf{X}_{\mathcal{B}_a} \mathbf{U}_{\mathcal{B}_b}] \delta}^n \right\} \leq \varepsilon. \quad (77)$$

- 2) For any nonempty disjoint pair of index sets $\mathcal{A}_b, \mathcal{B}_b \subseteq \mathcal{K}$,

$$\left| (g_{\mathcal{A}_a, \mathcal{A}_b}(\mathcal{M}_{\mathcal{A}_a}) \times g_{\mathcal{B}_a, \mathcal{B}_b}(\mathcal{M}_{\mathcal{B}_a})) \cap \mathcal{T}_{[\mathbf{U}_{\mathcal{A}_b} \mathbf{U}_{\mathcal{B}_b}] \delta}^n \right| \leq \exp(n(\mathbb{I}(\mathbf{U}_{\mathcal{A}_b} \mathbf{U}_{\mathcal{B}_b}; \mathbf{X}_{\mathcal{A}_b} \mathbf{X}_{\mathcal{B}_b}) + \varepsilon)) \quad (78)$$

for any index sets $\mathcal{A}_a \supseteq \mathcal{A}_b$, $\mathcal{B}_a \supseteq \mathcal{B}_b$ with $\mathcal{A}_a \cap \mathcal{B}_a = \emptyset$.

Consider $(\mu_\Omega, R_K) \in \mathcal{R}_i$ and choose \mathcal{U}_K as in Theorem 14. Fix $\varepsilon > 0$ and apply Lemma 20 to obtain encoding functions f_K and decoding functions $g_{\mathcal{A}_a, \mathcal{A}_b}$. For any pair $(\mathcal{A}, \mathcal{B}) \in \Omega$, find the nonempty subsets $\mathcal{A}_b \subseteq \mathcal{A}_a \subseteq \mathcal{A}$ and $\mathcal{B}_b \subseteq \mathcal{B}_a \subseteq \mathcal{B}$ such that (66)–(68) hold. (The case $\mathcal{A}_b = \emptyset$ or $\mathcal{B}_b = \emptyset$ can be ignored since due to (68) $\mathcal{A}_b = \emptyset$ or $\mathcal{B}_b = \emptyset$ it leads to $\mu_{\mathcal{A}, \mathcal{B}} \leq 0$, which is achieved by any code.) Define the functions $h_1 \triangleq g_{\mathcal{A}_a, \mathcal{A}_b} \circ f_{\mathcal{A}_a}$ and $h_2 \triangleq g_{\mathcal{B}_a, \mathcal{B}_b} \circ f_{\mathcal{B}_a}$. To analyze $\Theta(f_{\mathcal{A}}; f_{\mathcal{B}})$, we define $\mathcal{D}_1 \triangleq h_1(\mathcal{X}_{\mathcal{A}_a}^n)$ and partition $\mathcal{X}_{\mathcal{A}_a}^n$ as $\mathcal{X}_{\mathcal{A}_a}^n = \bigcup_{\mathbf{u}_{\mathcal{A}_b} \in \mathcal{D}_1} h_1^{-1}(\mathbf{u}_{\mathcal{A}_b})$. We may assume without loss of generality that $h_1^{-1}(\mathbf{u}_{\mathcal{A}_b}) \subseteq \mathcal{T}_{[\mathcal{X}_{\mathcal{A}_a} | \mathcal{U}_{\mathcal{A}_b}]^\delta}^n(\mathbf{u}_{\mathcal{A}_b})$ whenever $\mathbf{u}_{\mathcal{A}_b} \in \mathcal{D}_1 \cap \mathcal{T}_{[\mathcal{U}_{\mathcal{A}_b}]^\delta}^n$ as this does not interfere with the properties of the code. Defining \mathcal{D}_2 accordingly, we set $\mathcal{F} \triangleq (\mathcal{D}_1 \times \mathcal{D}_2) \cap \mathcal{T}_{[\mathcal{U}_{\mathcal{A}_b} \cup \mathcal{U}_{\mathcal{B}_b}]^\delta}^n$. Using the shorthand notation $\hat{\mathbf{U}}_1 \triangleq h_1(\mathbf{X}_{\mathcal{A}_a})$ and $\hat{\mathbf{U}}_2 \triangleq h_2(\mathbf{X}_{\mathcal{B}_a})$, we have

$$n \Theta(f_{\mathcal{A}}; f_{\mathcal{B}}) \stackrel{(a)}{\geq} n \Theta(f_{\mathcal{A}_a}; f_{\mathcal{B}_a}) \stackrel{(b)}{\geq} n \Theta(h_1; h_2) = I(h_1(\mathbf{X}_{\mathcal{A}_a}); h_2(\mathbf{X}_{\mathcal{B}_a})) \quad (79)$$

$$= \sum_{\mathbf{u}_{\mathcal{A}_b} \in \mathcal{D}_1, \mathbf{u}_{\mathcal{B}_b} \in \mathcal{D}_2} \mathbb{P}\{\hat{\mathbf{U}}_1 = \mathbf{u}_{\mathcal{A}_b}, \hat{\mathbf{U}}_2 = \mathbf{u}_{\mathcal{B}_b}\} \log \frac{\mathbb{P}\{\hat{\mathbf{U}}_1 = \mathbf{u}_{\mathcal{A}_b}, \hat{\mathbf{U}}_2 = \mathbf{u}_{\mathcal{B}_b}\}}{\mathbb{P}\{\hat{\mathbf{U}}_1 = \mathbf{u}_{\mathcal{A}_b}\} \mathbb{P}\{\hat{\mathbf{U}}_2 = \mathbf{u}_{\mathcal{B}_b}\}} \quad (80)$$

$$= \sum_{(\mathbf{u}_{\mathcal{A}_b}, \mathbf{u}_{\mathcal{B}_b}) \in \mathcal{F}} \mathbb{P}\{\hat{\mathbf{U}}_1 = \mathbf{u}_{\mathcal{A}_b}, \hat{\mathbf{U}}_2 = \mathbf{u}_{\mathcal{B}_b}\} \log \frac{\mathbb{P}\{\hat{\mathbf{U}}_1 = \mathbf{u}_{\mathcal{A}_b}, \hat{\mathbf{U}}_2 = \mathbf{u}_{\mathcal{B}_b}\}}{\mathbb{P}\{\hat{\mathbf{U}}_1 = \mathbf{u}_{\mathcal{A}_b}\} \mathbb{P}\{\hat{\mathbf{U}}_2 = \mathbf{u}_{\mathcal{B}_b}\}} \\ + \sum_{(\mathbf{u}_{\mathcal{A}_b}, \mathbf{u}_{\mathcal{B}_b}) \in \mathcal{F}^c} \mathbb{P}\{\hat{\mathbf{U}}_1 = \mathbf{u}_{\mathcal{A}_b}, \hat{\mathbf{U}}_2 = \mathbf{u}_{\mathcal{B}_b}\} \log \frac{\mathbb{P}\{\hat{\mathbf{U}}_1 = \mathbf{u}_{\mathcal{A}_b}, \hat{\mathbf{U}}_2 = \mathbf{u}_{\mathcal{B}_b}\}}{\mathbb{P}\{\hat{\mathbf{U}}_1 = \mathbf{u}_{\mathcal{A}_b}\} \mathbb{P}\{\hat{\mathbf{U}}_2 = \mathbf{u}_{\mathcal{B}_b}\}} \quad (81)$$

$$\stackrel{(c)}{\geq} p_{\mathcal{F}} \log \frac{p_{\mathcal{F}}}{\bar{p}_{\mathcal{F}}} + (1 - p_{\mathcal{F}}) \log \frac{1 - p_{\mathcal{F}}}{1 - \bar{p}_{\mathcal{F}}} \quad (82)$$

where (a) and (b) follow from the data processing inequality [30, Theorem 2.8.1] and (c) is a consequence of the log-sum inequality [30, Theorem 2.7.1]. Furthermore, we defined $p_{\mathcal{F}} \triangleq \mathbb{P}\{(\hat{\mathbf{U}}_1, \hat{\mathbf{U}}_2) \in \mathcal{F}\}$ and $\bar{p}_{\mathcal{F}} \triangleq \mathbb{P}\{(\bar{\mathbf{U}}_1, \bar{\mathbf{U}}_2) \in \mathcal{F}\}$ with $\bar{\mathbf{U}}_1 \triangleq h_1(\bar{\mathbf{X}}_{\mathcal{A}_a})$, $\bar{\mathbf{U}}_2 \triangleq h_2(\bar{\mathbf{X}}_{\mathcal{B}_a})$, where $(\bar{\mathbf{X}}_{\mathcal{A}_a}, \bar{\mathbf{X}}_{\mathcal{B}_a})$ are i.i.d. copies of $(\bar{\mathbf{X}}_{\mathcal{A}_a}, \bar{\mathbf{X}}_{\mathcal{B}_a}) \sim p_{\mathbf{X}_{\mathcal{A}_a}} p_{\mathbf{X}_{\mathcal{B}_a}}$. The expression (82) can be further bounded as

$$p_{\mathcal{F}} \log \frac{p_{\mathcal{F}}}{\bar{p}_{\mathcal{F}}} + (1 - p_{\mathcal{F}}) \log \frac{1 - p_{\mathcal{F}}}{1 - \bar{p}_{\mathcal{F}}} = -h_0(p_{\mathcal{F}}) - p_{\mathcal{F}} \log \bar{p}_{\mathcal{F}} - (1 - p_{\mathcal{F}}) \log(1 - \bar{p}_{\mathcal{F}}) \quad (83)$$

$$\geq -h_0(p_{\mathcal{F}}) - p_{\mathcal{F}} \log \bar{p}_{\mathcal{F}} \stackrel{(d)}{\geq} -h_0(p_{\mathcal{F}}) - (1 - \varepsilon) \log \bar{p}_{\mathcal{F}} \quad (84)$$

$$\geq -\log(2) - (1 - \varepsilon) \log \bar{p}_{\mathcal{F}}, \quad (85)$$

where (d) follows from (77). For each $\mathbf{u}_{\mathcal{A}_b} \in \mathcal{D}_1$ and $\mathbf{u}_{\mathcal{B}_b} \in \mathcal{D}_2$ define

$$\mathcal{S}(\mathbf{u}_{\mathcal{A}_b}, \mathbf{u}_{\mathcal{B}_b}) \triangleq \{\mathbf{u}_{\mathcal{A}_b}\} \times h_1^{-1}(\mathbf{u}_{\mathcal{A}_b}) \times h_2^{-1}(\mathbf{u}_{\mathcal{B}_b}) \times \{\mathbf{u}_{\mathcal{B}_b}\} \quad (86)$$

and

$$\mathbb{S} \triangleq \bigcup_{(\mathbf{u}_{\mathcal{A}_b}, \mathbf{u}_{\mathcal{B}_b}) \in \mathcal{F}} \mathcal{S}(\mathbf{u}_{\mathcal{A}_b}, \mathbf{u}_{\mathcal{B}_b}). \quad (87)$$

Now, pick any $(\hat{\mathbf{u}}_{\mathcal{A}_b}, \hat{\mathbf{x}}_{\mathcal{A}_a}, \hat{\mathbf{x}}_{\mathcal{B}_a}, \hat{\mathbf{u}}_{\mathcal{B}_b}) \in \mathbb{S}$. Let $\hat{\mathbf{U}}_{\mathcal{A}_b}$, $\hat{\mathbf{X}}_{\mathcal{A}_a}$, $\hat{\mathbf{X}}_{\mathcal{B}_a}$, and $\hat{\mathbf{U}}_{\mathcal{B}_b}$ be the type variables corresponding to $\hat{\mathbf{u}}_{\mathcal{A}_b}$, $\hat{\mathbf{x}}_{\mathcal{A}_a}$, $\hat{\mathbf{x}}_{\mathcal{B}_a}$, and $\hat{\mathbf{u}}_{\mathcal{B}_b}$, respectively. From part 1 of Lemma 23 we know

$$\mathbb{P}\{\bar{\mathbf{X}}_{\mathcal{A}_a} = \hat{\mathbf{x}}_{\mathcal{A}_a}, \bar{\mathbf{X}}_{\mathcal{B}_a} = \hat{\mathbf{x}}_{\mathcal{B}_a}\} = \exp(-n(H(\hat{\mathbf{X}}_{\mathcal{A}_a} \hat{\mathbf{X}}_{\mathcal{B}_a}) + D_{\text{KL}}(\hat{\mathbf{X}}_{\mathcal{A}_a} \hat{\mathbf{X}}_{\mathcal{B}_a} \| \bar{\mathbf{X}}_{\mathcal{A}_a} \bar{\mathbf{X}}_{\mathcal{B}_a}))). \quad (88)$$

Let $\kappa(\mathbf{u}_{\mathcal{A}_b}, \mathbf{u}_{\mathcal{B}_b}; \hat{\mathbf{U}}_{\mathcal{A}_b}, \hat{\mathbf{X}}_{\mathcal{A}_a}, \hat{\mathbf{X}}_{\mathcal{B}_a}, \hat{\mathbf{U}}_{\mathcal{B}_b})$ denote the number of elements in $\mathcal{S}(\mathbf{u}_{\mathcal{A}_b}, \mathbf{u}_{\mathcal{B}_b})$ with type $(\hat{\mathbf{U}}_{\mathcal{A}_b}, \hat{\mathbf{X}}_{\mathcal{A}_a}, \hat{\mathbf{X}}_{\mathcal{B}_a}, \hat{\mathbf{U}}_{\mathcal{B}_b})$. Then, by part 2 of Lemma 23

$$\kappa(\mathbf{u}_{\mathcal{A}_b}, \mathbf{u}_{\mathcal{B}_b}; \hat{\mathbf{U}}_{\mathcal{A}_b}, \hat{\mathbf{X}}_{\mathcal{A}_a}, \hat{\mathbf{X}}_{\mathcal{B}_a}, \hat{\mathbf{U}}_{\mathcal{B}_b}) \leq \exp(nH(\hat{\mathbf{X}}_{\mathcal{A}_a} \hat{\mathbf{X}}_{\mathcal{B}_a} | \hat{\mathbf{U}}_{\mathcal{A}_b} \hat{\mathbf{U}}_{\mathcal{B}_b})). \quad (89)$$

Letting $\kappa(\hat{U}_{\mathcal{A}_b}, \hat{X}_{\mathcal{A}_a}, \hat{X}_{\mathcal{B}_a}, \hat{U}_{\mathcal{B}_b})$ be the number of elements of \mathbb{S} with type $(\hat{U}_{\mathcal{A}_b}, \hat{X}_{\mathcal{A}_a}, \hat{X}_{\mathcal{B}_a}, \hat{U}_{\mathcal{B}_b})$, we have

$$\kappa(\hat{U}_{\mathcal{A}_b}, \hat{X}_{\mathcal{A}_a}, \hat{X}_{\mathcal{B}_a}, \hat{U}_{\mathcal{B}_b}) = \sum_{(\mathbf{u}_{\mathcal{A}_b}, \mathbf{u}_{\mathcal{B}_b}) \in \mathcal{F}} \kappa(\mathbf{u}_{\mathcal{A}_b}, \mathbf{u}_{\mathcal{B}_b}; \hat{U}_{\mathcal{A}_b}, \hat{X}_{\mathcal{A}_a}, \hat{X}_{\mathcal{B}_a}, \hat{U}_{\mathcal{B}_b}) \quad (90)$$

$$\stackrel{(a)}{\leq} \sum_{(\mathbf{u}_{\mathcal{A}_b}, \mathbf{u}_{\mathcal{B}_b}) \in \mathcal{F}} \exp(nH(\hat{X}_{\mathcal{A}_a} \hat{X}_{\mathcal{B}_a} | \hat{U}_{\mathcal{A}_b} \hat{U}_{\mathcal{B}_b})) \quad (91)$$

$$\stackrel{(b)}{\leq} \exp(n(I(\mathbf{U}_{\mathcal{A}_b} \mathbf{U}_{\mathcal{B}_b}; \mathbf{X}_{\mathcal{A}_a} \mathbf{X}_{\mathcal{B}_a}) + H(\hat{X}_{\mathcal{A}_a} \hat{X}_{\mathcal{B}_a} | \hat{U}_{\mathcal{A}_b} \hat{U}_{\mathcal{B}_b}) + \varepsilon)), \quad (92)$$

where (a) follows from (89) and (b) from (78). Thus,

$$\begin{aligned} \mathbb{P}\{(\bar{\mathbf{U}}_{\mathcal{A}_b}, \bar{\mathbf{U}}_{\mathcal{B}_b}) \in \mathcal{F}\} &\stackrel{(a)}{=} \sum_{\hat{U}_{\mathcal{A}_b}, \hat{X}_{\mathcal{A}_a}, \hat{X}_{\mathcal{B}_a}, \hat{U}_{\mathcal{B}_b}} \kappa(\hat{U}_{\mathcal{A}_b}, \hat{X}_{\mathcal{A}_a}, \hat{X}_{\mathcal{B}_a}, \hat{U}_{\mathcal{B}_b}) \\ &\quad \cdot \exp\left(-n(H(\hat{X}_{\mathcal{A}_a} \hat{X}_{\mathcal{B}_a}) + D_{\text{KL}}(\hat{X}_{\mathcal{A}_a} \hat{X}_{\mathcal{B}_a} \| \bar{\mathbf{X}}_{\mathcal{A}_a} \bar{\mathbf{X}}_{\mathcal{B}_a}))\right) \end{aligned} \quad (93)$$

$$\stackrel{(b)}{\leq} \sum_{\hat{U}_{\mathcal{A}_b}, \hat{X}_{\mathcal{A}_a}, \hat{X}_{\mathcal{B}_a}, \hat{U}_{\mathcal{B}_b}} \exp\left(-n(k(\hat{U}_{\mathcal{A}_b}, \hat{X}_{\mathcal{A}_a}, \hat{X}_{\mathcal{B}_a}, \hat{U}_{\mathcal{B}_b}) - \varepsilon)\right), \quad (94)$$

where the sum is over all types that occur in \mathbb{S} . Here, (a) follows from (88) and (b) from (92) and we defined

$$\begin{aligned} k(\hat{U}_{\mathcal{A}_b}, \hat{X}_{\mathcal{A}_a}, \hat{X}_{\mathcal{B}_a}, \hat{U}_{\mathcal{B}_b}) &\triangleq I(\hat{U}_{\mathcal{A}_b} \hat{U}_{\mathcal{B}_b}; \hat{X}_{\mathcal{A}_a} \hat{X}_{\mathcal{B}_a}) - I(\mathbf{U}_{\mathcal{A}_b} \mathbf{U}_{\mathcal{B}_b}; \mathbf{X}_{\mathcal{A}_a} \mathbf{X}_{\mathcal{B}_a}) \\ &\quad + D_{\text{KL}}(\hat{X}_{\mathcal{A}_a} \hat{X}_{\mathcal{B}_a} \| \bar{\mathbf{X}}_{\mathcal{A}_a} \bar{\mathbf{X}}_{\mathcal{B}_a}). \end{aligned} \quad (95)$$

Using a type counting argument (Lemma 22) we can further bound

$$\begin{aligned} \mathbb{P}\{(\bar{\mathbf{U}}_{\mathcal{A}_b}, \bar{\mathbf{U}}_{\mathcal{B}_b}) \in \mathcal{F}\} &\leq (n+1)^{|\mathcal{U}_{\mathcal{A}_b}| |\mathcal{X}_{\mathcal{A}_a}| |\mathcal{X}_{\mathcal{B}_a}| |\mathcal{U}_{\mathcal{B}_b}|} \\ &\quad \cdot \max_{\hat{U}_{\mathcal{A}_b}, \hat{X}_{\mathcal{A}_a}, \hat{X}_{\mathcal{B}_a}, \hat{U}_{\mathcal{B}_b}} \exp\left(-n(k(\hat{U}_{\mathcal{A}_b}, \hat{X}_{\mathcal{A}_a}, \hat{X}_{\mathcal{B}_a}, \hat{U}_{\mathcal{B}_b}) - \varepsilon)\right), \end{aligned} \quad (96)$$

where the maximum is over all types occurring in \mathbb{S} . For any type $(\hat{U}_{\mathcal{A}_b}, \hat{X}_{\mathcal{A}_a}, \hat{X}_{\mathcal{B}_a}, \hat{U}_{\mathcal{B}_b})$ in \mathbb{S} , we have by construction $(\hat{U}_{\mathcal{A}_b}, \hat{X}_{\mathcal{A}_a}, \hat{X}_{\mathcal{B}_a}, \hat{U}_{\mathcal{B}_b}) \in \mathcal{L}_\delta(\mathbf{U}_{\mathcal{A}_b}, \mathbf{X}_{\mathcal{A}_a}, \mathbf{X}_{\mathcal{B}_a}, \mathbf{U}_{\mathcal{B}_b})$ (recall Definition 10) and we can thus conclude

$$\begin{aligned} \mathbb{P}\{(\bar{\mathbf{U}}_{\mathcal{A}_b}, \bar{\mathbf{U}}_{\mathcal{B}_b}) \in \mathcal{F}\} &\leq (n+1)^{|\mathcal{U}_{\mathcal{K}}| |\mathcal{X}_{\mathcal{K}}|} \\ &\quad \cdot \max_{(\tilde{\mathbf{U}}_{\mathcal{A}_b}, \tilde{\mathbf{X}}_{\mathcal{A}_a}, \tilde{\mathbf{X}}_{\mathcal{B}_a}, \tilde{\mathbf{U}}_{\mathcal{B}_b}) \in \mathcal{L}_\delta(\mathbf{U}_{\mathcal{A}_b}, \mathbf{X}_{\mathcal{A}_a}, \mathbf{X}_{\mathcal{B}_a}, \mathbf{U}_{\mathcal{B}_b})} \exp\left(-n(k(\tilde{\mathbf{U}}_{\mathcal{A}_b}, \tilde{\mathbf{X}}_{\mathcal{A}_a}, \tilde{\mathbf{X}}_{\mathcal{B}_a}, \tilde{\mathbf{U}}_{\mathcal{B}_b}) - \varepsilon)\right). \end{aligned} \quad (97)$$

Combining (85) and (97) we showed that for n large enough

$$\Theta(f_{\mathcal{A}}; f_{\mathcal{B}}) \geq -\frac{\log(2)}{n} - \frac{1-\varepsilon}{n} \log \mathbb{P}\{(\bar{\mathbf{U}}_{\mathcal{A}_b}, \bar{\mathbf{U}}_{\mathcal{B}_b}) \in \mathcal{F}\} \quad (98)$$

$$\begin{aligned} &\geq -\varepsilon + (1-\varepsilon) \\ &\quad \cdot \min_{(\tilde{\mathbf{U}}_{\mathcal{A}_b}, \tilde{\mathbf{X}}_{\mathcal{A}_a}, \tilde{\mathbf{X}}_{\mathcal{B}_a}, \tilde{\mathbf{U}}_{\mathcal{B}_b}) \in \mathcal{L}_\delta(\mathbf{U}_{\mathcal{A}_b}, \mathbf{X}_{\mathcal{A}_a}, \mathbf{X}_{\mathcal{B}_a}, \mathbf{U}_{\mathcal{B}_b})} \left(k(\tilde{\mathbf{U}}_{\mathcal{A}_b}, \tilde{\mathbf{X}}_{\mathcal{A}_a}, \tilde{\mathbf{X}}_{\mathcal{B}_a}, \tilde{\mathbf{U}}_{\mathcal{B}_b}) - \varepsilon\right) \end{aligned} \quad (99)$$

$$\geq -2\varepsilon + (1-\varepsilon) \min_{(\tilde{\mathbf{U}}_{\mathcal{A}_b}, \tilde{\mathbf{X}}_{\mathcal{A}_a}, \tilde{\mathbf{X}}_{\mathcal{B}_a}, \tilde{\mathbf{U}}_{\mathcal{B}_b}) \in \mathcal{L}_\delta(\mathbf{U}_{\mathcal{A}_b}, \mathbf{X}_{\mathcal{A}_a}, \mathbf{X}_{\mathcal{B}_a}, \mathbf{U}_{\mathcal{B}_b})} k(\tilde{\mathbf{U}}_{\mathcal{A}_b}, \tilde{\mathbf{X}}_{\mathcal{A}_a}, \tilde{\mathbf{X}}_{\mathcal{B}_a}, \tilde{\mathbf{U}}_{\mathcal{B}_b}) \quad (100)$$

$$\geq \min_{(\tilde{\mathbf{U}}_{\mathcal{A}_b}, \tilde{\mathbf{X}}_{\mathcal{A}_a}, \tilde{\mathbf{X}}_{\mathcal{B}_a}, \tilde{\mathbf{U}}_{\mathcal{B}_b}) \in \mathcal{L}_\delta(\mathbf{U}_{\mathcal{A}_b}, \mathbf{X}_{\mathcal{A}_a}, \mathbf{X}_{\mathcal{B}_a}, \mathbf{U}_{\mathcal{B}_b})} k(\tilde{\mathbf{U}}_{\mathcal{A}_b}, \tilde{\mathbf{X}}_{\mathcal{A}_a}, \tilde{\mathbf{X}}_{\mathcal{B}_a}, \tilde{\mathbf{U}}_{\mathcal{B}_b}) - (2 + I(\mathbf{X}_{\mathcal{A}_a}; \mathbf{X}_{\mathcal{B}_a}))\varepsilon \quad (101)$$

$$= \min_{(\tilde{\mathbf{U}}_{\mathcal{A}_b}, \tilde{\mathbf{X}}_{\mathcal{A}_a}, \tilde{\mathbf{X}}_{\mathcal{B}_a}, \tilde{\mathbf{U}}_{\mathcal{B}_b}) \in \mathcal{L}_\delta(\mathbf{U}_{\mathcal{A}_b}, \mathbf{X}_{\mathcal{A}_a}, \mathbf{X}_{\mathcal{B}_a}, \mathbf{U}_{\mathcal{B}_b})} k(\tilde{\mathbf{U}}_{\mathcal{A}_b}, \tilde{\mathbf{X}}_{\mathcal{A}_a}, \tilde{\mathbf{X}}_{\mathcal{B}_a}, \tilde{\mathbf{U}}_{\mathcal{B}_b}) - C\varepsilon \quad (102)$$

for some constant C . As $k(\tilde{\mathbf{U}}_{\mathcal{A}_b}, \tilde{\mathbf{X}}_{\mathcal{A}_a}, \tilde{\mathbf{X}}_{\mathcal{B}_a}, \tilde{\mathbf{U}}_{\mathcal{B}_b})$ is continuous as a function of $p_{\tilde{\mathbf{U}}_{\mathcal{A}_b}, \tilde{\mathbf{X}}_{\mathcal{A}_a}, \tilde{\mathbf{X}}_{\mathcal{B}_a}, \tilde{\mathbf{U}}_{\mathcal{B}_b}}$ and (102) holds for arbitrarily small δ , it follows that for large enough n

$$\Theta(f_{\mathcal{A}}; f_{\mathcal{B}}) \geq \min_{(\tilde{\mathbf{U}}_{\mathcal{A}_b}, \tilde{\mathbf{X}}_{\mathcal{A}_a}, \tilde{\mathbf{X}}_{\mathcal{B}_a}, \tilde{\mathbf{U}}_{\mathcal{B}_b}) \in \mathcal{L}(\mathbf{U}_{\mathcal{A}_b}, \mathbf{X}_{\mathcal{A}_a}, \mathbf{X}_{\mathcal{B}_a}, \mathbf{U}_{\mathcal{B}_b})} k(\tilde{\mathbf{U}}_{\mathcal{A}_b}, \tilde{\mathbf{X}}_{\mathcal{A}_a}, \tilde{\mathbf{X}}_{\mathcal{B}_a}, \tilde{\mathbf{U}}_{\mathcal{B}_b}) - C'\varepsilon \quad (103)$$

for some (larger) constant C' by letting $\delta \rightarrow 0$. As in the proof of Theorem 5, observe that for $(\tilde{U}_{\mathcal{A}_b}, \tilde{X}_{\mathcal{A}_a}, \tilde{X}_{\mathcal{B}_a}, \tilde{U}_{\mathcal{B}_b}) \in \mathcal{L}(\mathcal{U}_{\mathcal{A}_b}, \mathcal{X}_{\mathcal{A}_a}, \mathcal{X}_{\mathcal{B}_a}, \mathcal{U}_{\mathcal{B}_b})$ we have

$$k(\tilde{U}_{\mathcal{A}_b}, \tilde{X}_{\mathcal{A}_a}, \tilde{X}_{\mathcal{B}_a}, \tilde{U}_{\mathcal{B}_b}) = I(\tilde{U}_{\mathcal{A}_b} \tilde{X}_{\mathcal{A}_a}; \tilde{X}_{\mathcal{B}_a} \tilde{U}_{\mathcal{B}_b}). \quad (104)$$

Like in the proof of Theorem 5, one can also show

$$\min_{(\tilde{U}_{\mathcal{A}_b}, \tilde{X}_{\mathcal{A}_a}, \tilde{X}_{\mathcal{B}_a}, \tilde{U}_{\mathcal{B}_b}) \in \mathcal{L}(\mathcal{U}_{\mathcal{A}_b}, \mathcal{X}_{\mathcal{A}_a}, \mathcal{X}_{\mathcal{B}_a}, \mathcal{U}_{\mathcal{B}_b})} I(\tilde{U}_{\mathcal{A}_b} \tilde{X}_{\mathcal{A}_a}; \tilde{X}_{\mathcal{B}_a} \tilde{U}_{\mathcal{B}_b}) \geq I(\mathcal{U}_{\mathcal{A}_b}; \mathcal{U}_{\mathcal{B}_b}). \quad (105)$$

Combining (103)–(105), we have

$$\Theta(f_{\mathcal{A}}; f_{\mathcal{B}}) \geq I(\mathcal{U}_{\mathcal{A}_b}; \mathcal{U}_{\mathcal{B}_b}) - C' \varepsilon \stackrel{(a)}{\geq} \mu_{\mathcal{A}, \mathcal{B}} - C' \varepsilon, \quad (106)$$

where (a) follows from assumption (68). We hence obtain $(\mu_{2^{\mathcal{A}}, 2^{\mathcal{B}}} - C' \varepsilon, R_{\mathcal{K}} + \varepsilon) \in \mathcal{R}$; since ε was arbitrary, this completes the proof.

IV. SUMMARY AND DISCUSSION

We introduced a novel multi-terminal source coding problem termed information-theoretic biclustering. Interestingly, this problem is related to several other problems at the frontier of statistics and information theory and offers a formidable mathematical complexity. Indeed, it is fundamentally different from “classical” distributed source coding problems where the encoders usually aim at reducing, as much as possible, redundant information among the sources while still satisfying a fidelity criterion. Whereas in the considered problem, the encoders are interested in maximizing precisely such redundant information.

Although an exact characterization of the achievable region is mathematically very challenging and still remains elusive, we provided tight outer and inner bounds to the set of achievable rates. We thoroughly studied the special case of two symmetric binary sources for which novel cardinality bounding techniques were developed. Based on numerical evidence we formulated a conjecture that entails an explicit expression for the inner bound. This conjecture provides strong evidence that our inner and outer bound do not meet in general. We firmly believe that an improved outer bound, satisfying the adequate Markov chains, is required for a tight characterization of the achievable region.

We further established analogous bounds to the achievable rate region of information-theoretic biclustering with more than two sources. However, these bounds cannot be tight since the infamous Körner-Marton problem constitutes a counterexample. For an analogue of the well-known CEO problem we showed that our bounds are tight in a special case, leveraging existing results from multiterminal lossy source coding.

The interesting challenge of the biclustering problem lies in the fact that one needs to bound the mutual information between two arbitrary encodings solely based on their rates. Available information-theoretic manipulations seem incapable of handling this requirement well.

APPENDIX

A. Types, Typical Sequences and Related Results

In this appendix we introduce standard notion and results, as needed for the mathematical developments and proofs in this work. The results can be easily derived from the standard formulations provided in [10] and [33].

Definition 21 (Type; [33, Definition 2.1]). *The type of a vector $\mathbf{x} \in \mathcal{X}^n$ is the random variable $\hat{X} \sim p_{\hat{X}} \in \mathcal{P}(\mathcal{X})$ defined by*

$$p_{\hat{X}}(x) = \frac{1}{n} N(x|\mathbf{x}), \text{ for every } x \in \mathcal{X}, \quad (107)$$

where $N(x|\mathbf{x})$ denotes the number of occurrences of x in \mathbf{x} . For a random variable \hat{X} , the set of n -sequences with type \hat{X} is denoted $\mathcal{T}_{\hat{X}}^n$.

For a pair of random variables (X, Y) , we say that $\mathbf{y} \in \mathcal{Y}^n$ has conditional type Y given $\mathbf{x} \in \mathcal{X}^n$ if $(\mathbf{x}, \mathbf{y}) \in \mathcal{T}_{XY}^n$. The set of all n -sequences $\mathbf{y} \in \mathcal{Y}^n$ with conditional type Y given \mathbf{x} will be denoted $\mathcal{T}_{Y|\mathbf{x}}^n$.

A key property of types is the following result, known as type counting.

Lemma 22 (Type counting; [33, Lemma 2.2]). *The number of different types of sequences in \mathcal{X}^n is less than $(n+1)^{|\mathcal{X}|}$.*

Some important properties of types are listed in the following lemma.

Lemma 23 ([33, Lemmas 2.5 and 2.6]).

- 1) *For any two random variables X, \tilde{X} on \mathcal{X} , and $\mathbf{x} \in \mathcal{T}_X^n$*

$$P\{\tilde{\mathbf{X}} = \mathbf{x}\} = \exp(-n(H(X) + D_{\text{KL}}(X\|\tilde{X}))), \quad (108)$$

where $\tilde{\mathbf{X}}$ is a sequence of n i.i.d. copies of \tilde{X} .

- 2) *For a pair of random variables (X, Y) and $\mathbf{x} \in \mathcal{X}^n$, such that $\mathcal{T}_{Y|X}^n(\mathbf{x}) \neq \emptyset$*

$$(n+1)^{-|\mathcal{X}||\mathcal{Y}|} \exp(nH(Y|X)) \leq |\mathcal{T}_{Y|X}^n(\mathbf{x})| \leq \exp(nH(Y|X)). \quad (109)$$

Definition 24 (Typicality; [10, Section 2.4]). *Consider $X \sim p_X \in \mathcal{P}(\mathcal{X})$ and $\delta \geq 0$. We call the random variable $Y \sim p_Y \in \mathcal{P}(\mathcal{Y})$ δ -typical if $Y \in \mathcal{T}_{[X]\delta}$ with*

$$\mathcal{T}_{[X]\delta} \triangleq \{\tilde{X} \sim p_{\tilde{X}} \in \mathcal{P}(\mathcal{X}) : |p_{\tilde{X}}(x) - p_X(x)| \leq \delta p_X(x), \forall x \in \mathcal{X}\}. \quad (110)$$

A sequence $\mathbf{x} \in \mathcal{X}^n$ is δ -typical if its type \hat{X} is δ -typical. The set of all δ -typical n -sequences is denoted $\mathcal{T}_{[X]\delta}^n$.

Note that Y is 0-typical if and only if $Y \simeq X$. Given $p_{X,Y} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ we call the elements of $\mathcal{T}_{[X]Y\delta}^n$ the *jointly δ -typical n -sequences*. We also define the *conditionally typical n -sequences* $\mathcal{T}_{[Y|X]\delta}^n(\mathbf{x}) \triangleq \{\mathbf{y} \in \mathcal{Y}^n : (\mathbf{x}, \mathbf{y}) \in \mathcal{T}_{[X]Y\delta}^n\}$. Typical sequences have several useful properties, which are presented in the following.

Lemma 25 (Size of typical sets; [10, Sections 2.4 and 2.5]). *The following properties hold for $(X, Y) \sim p_{X,Y}$:*

- 1) *Using $\varepsilon(\delta) = \delta H(X)$,*

$$|\mathcal{T}_{[X]\delta}^n| \leq e^{n(H(X) + \varepsilon(\delta))}. \quad (111)$$

- 2) *For $\delta > 0$, $\varepsilon' > 0$, n sufficiently large (as a function of ε' and p_X), and $\varepsilon(\delta) = \delta H(X)$,*

$$|\mathcal{T}_{[X]\delta}^n| \geq (1 - \varepsilon') e^{n(H(X) - \varepsilon(\delta))}. \quad (112)$$

- 3) *For $\mathbf{x} \in \mathcal{X}^n$ and $\varepsilon(\delta) = \delta H(Y|X)$,*

$$|\mathcal{T}_{[Y|X]\delta}^n(\mathbf{x})| \leq e^{n(H(Y|X) + \varepsilon(\delta))}. \quad (113)$$

- 4) *Let $\delta' > 0$ and $\mathbf{x} \in \mathcal{T}_{[X]\delta'}^n$. For $\delta > \delta'$, $\varepsilon' > 0$, n sufficiently large (as a function of ε' and $p_{X,Y}$), and $\varepsilon(\delta) = \delta H(Y|X)$,*

$$|\mathcal{T}_{[Y|X]\delta}^n(\mathbf{x})| \geq (1 - \varepsilon') e^{n(H(Y|X) - \varepsilon(\delta))}. \quad (114)$$

Lemma 26 (Generalized Markov Lemma; [37, Lemma 3.4]). *Let $X_{\mathcal{K}}$ and $U_{\mathcal{K}}$ be such that $U_k \ominus X_k \ominus (X_{\mathcal{K} \setminus k}, U_{\mathcal{K} \setminus k})$ and fix $\varepsilon > 0$. Choose a sufficiently small $\delta > 0$ and for each $k \in \mathcal{K}$, let $\tilde{\mathbf{U}}_{k, \mathcal{M}_k}$ ($\mathcal{M}_k \triangleq \{1, 2, \dots, M_k\}$) be a sequence of mutually independent random variables, also independent of $X_{\mathcal{K}}$ and $U_{\mathcal{K}}$, taking values equiprobably in $\mathcal{T}_{[U_k]\delta}^n$. Assume $M_k > \exp(nI(X_k; U_k))$. Then, for sufficiently large n there exist functions $f_k: \mathcal{X}_{\mathcal{K}}^n \times (\mathcal{U}_{\mathcal{K}}^n)^{M_k} \rightarrow \{1, 2, \dots, M_k\}$ such that, using $W_k = f_k(\mathbf{X}_k, \tilde{\mathbf{U}}_{k, \mathcal{M}_k})$ and $\mathbf{U}_k^* = \tilde{\mathbf{U}}_{k, W_k}$*

$$P\{(\mathbf{X}_{\mathcal{K}}, \mathbf{U}_{\mathcal{K}}^*) \in \mathcal{T}_{[X_{\mathcal{K}} U_{\mathcal{K}}]\delta}^n\} \geq 1 - \varepsilon. \quad (115)$$

B. Proof of Theorem 3

For $(\mu, R_1, R_2) \in \mathcal{R}$, let (f, g) be an (n, R_1, R_2) code for some $n \in \mathbb{N}$ such that $\Theta(f; g) \geq \mu$. We define the random variables $U_l \triangleq (\mathbf{X}^{l-1}, f(\mathbf{X}))$ and $V_l \triangleq (\mathbf{Z}^{l-1}, g(\mathbf{Z}))$ and obtain

$$nR_1 \geq H(f(\mathbf{X})) = I(f(\mathbf{X}); \mathbf{X}) \quad (116)$$

$$= \sum_{l=1}^n I(\mathbf{x}_l; f(\mathbf{x}) | \mathbf{x}^{l-1}) \quad (117)$$

$$= \sum_{l=1}^n I(\mathbf{x}_l; \mathbf{U}_l) \quad (118)$$

and accordingly

$$nR_2 \geq \sum_{l=1}^n I(\mathbf{Z}_l; \mathbf{V}_l). \quad (119)$$

We also have

$$n\mu \leq I(f(\mathbf{X}); g(\mathbf{Z})) \quad (120)$$

$$= I(f(\mathbf{X}); \mathbf{X}) - I(f(\mathbf{X}); \mathbf{X} | g(\mathbf{Z})) \quad (121)$$

$$= I(f(\mathbf{X}); \mathbf{X}) + I(g(\mathbf{Z}); \mathbf{Z}) - I(f(\mathbf{X}); \mathbf{X} | g(\mathbf{Z})) - I(g(\mathbf{Z}); \mathbf{Z}) \quad (122)$$

$$= I(f(\mathbf{X}); \mathbf{X}) + I(g(\mathbf{Z}); \mathbf{Z}) - I(f(\mathbf{X}); \mathbf{XZ} | g(\mathbf{Z})) - I(g(\mathbf{Z}); \mathbf{XZ}) \quad (123)$$

$$= I(f(\mathbf{X}); \mathbf{X}) + I(g(\mathbf{Z}); \mathbf{Z}) - I(f(\mathbf{X}), g(\mathbf{Z}); \mathbf{XZ}) \quad (124)$$

$$= \sum_{l=1}^n [I(\mathbf{U}_l; \mathbf{x}_l) + I(\mathbf{V}_l; \mathbf{Z}_l) - I(\mathbf{U}_l \mathbf{V}_l; \mathbf{x}_l \mathbf{Z}_l)]. \quad (125)$$

Now a standard time-sharing argument shows $\mathcal{R} \subseteq \mathcal{R}_o$. To see $\mathcal{R}_o \subseteq \mathcal{R}'_o$, note that

$$I(\mathbf{U}; \mathbf{X}) + I(\mathbf{V}; \mathbf{Z}) - I(\mathbf{UV}; \mathbf{XZ}) \quad (126)$$

$$= I(\mathbf{V}; \mathbf{Z}) - I(\mathbf{V}; \mathbf{XZ} | \mathbf{U}) \quad (127)$$

$$= I(\mathbf{U}; \mathbf{Z}) + I(\mathbf{V}; \mathbf{Z}) - I(\mathbf{U}; \mathbf{Z}) - I(\mathbf{V}; \mathbf{XZ} | \mathbf{U}) \quad (128)$$

$$= I(\mathbf{U}; \mathbf{Z}) + I(\mathbf{V}; \mathbf{Z}) - I(\mathbf{U}; \mathbf{Z}) - I(\mathbf{V}; \mathbf{Z} | \mathbf{U}) - I(\mathbf{V}; \mathbf{X} | \mathbf{ZU}) \quad (129)$$

$$= I(\mathbf{U}; \mathbf{Z}) + I(\mathbf{V}; \mathbf{Z}) - I(\mathbf{UV}; \mathbf{Z}) - I(\mathbf{V}; \mathbf{X} | \mathbf{ZU}) \quad (130)$$

$$= I(\mathbf{U}; \mathbf{Z}) - I(\mathbf{U}; \mathbf{Z} | \mathbf{V}) - I(\mathbf{V}; \mathbf{X} | \mathbf{ZU}) \quad (131)$$

$$\leq I(\mathbf{U}; \mathbf{Z}) \quad (132)$$

and by a symmetric argument

$$I(\mathbf{U}; \mathbf{X}) + I(\mathbf{V}; \mathbf{Z}) - I(\mathbf{UV}; \mathbf{XZ}) \leq I(\mathbf{V}; \mathbf{X}). \quad (133)$$

C. Proof of Proposition 4

Most steps in the proof apply to both \mathcal{R}_o and \mathcal{R}'_o . We thus state the proof for \mathcal{R}_o and point out the required modifications where appropriate.

Define the set of pmfs (with finite, but arbitrarily large support)

$$\mathcal{Q} \triangleq \{p_{\mathbf{U}, \mathbf{V}, \mathbf{X}, \mathbf{Z}} : \mathbf{X}' \text{---} \mathbf{Z}' \text{---} \mathbf{V}', \mathbf{U}' \text{---} \mathbf{X}' \text{---} \mathbf{Z}', \text{ and } (\mathbf{X}', \mathbf{Z}') \sim (\mathbf{X}, \mathbf{Z})\} \quad (134)$$

and the compact set of pmfs with fixed alphabet size

$$\mathcal{Q}(a, b) \triangleq \{p_{\mathbf{U}, \mathbf{V}, \mathbf{X}, \mathbf{Z}} \in \mathcal{Q} : |\mathcal{U}'| = a, |\mathcal{V}'| = b\}. \quad (135)$$

Define the continuous vector valued function $\mathbf{F} \triangleq (F_1, F_2, F_3)$ as

$$F_1(p_{\mathbf{U}, \mathbf{V}, \mathbf{X}, \mathbf{Z}}) \triangleq I(\mathbf{X}; \mathbf{U}) + I(\mathbf{Z}; \mathbf{V}) - I(\mathbf{UV}; \mathbf{XZ}), \quad (136)$$

$$F_2(p_{\mathbf{U}, \mathbf{V}, \mathbf{X}, \mathbf{Z}}) \triangleq I(\mathbf{U}; \mathbf{X}), \quad (137)$$

$$F_3(p_{\mathbf{U}, \mathbf{V}, \mathbf{X}, \mathbf{Z}}) \triangleq I(\mathbf{V}; \mathbf{Z}). \quad (138)$$

In the proof of \mathcal{R}'_o , the definition of F_1 is replaced with $F_1(p_{\mathbf{U}, \mathbf{V}, \mathbf{X}, \mathbf{Z}}) \triangleq \min(I(\mathbf{U}; \mathbf{Z}), I(\mathbf{V}; \mathbf{X}))$. We can now write

$\mathcal{R}_o = \mathbf{F}(\mathcal{Q}) + \mathcal{O}$ and $\mathcal{S}_o = \mathbf{F}(\mathcal{Q}(|\mathcal{X}|, |\mathcal{Z}|)) + \mathcal{O}$ where $\mathcal{O} \triangleq (\mathbb{R}_- \times \mathbb{R}_+ \times \mathbb{R}_+)$. Thus, we need to show

$$\mathcal{R}_o = \text{conv}(\mathbf{F}(\mathcal{Q}(|\mathcal{X}|, |\mathcal{Z}|)) + \mathcal{O}). \quad (139)$$

Since \mathcal{R}_o is convex, by defining the extended real function $\tilde{\psi}(\boldsymbol{\lambda}) \triangleq \sup_{\mathbf{x} \in \mathcal{R}_o} \boldsymbol{\lambda} \cdot \mathbf{x}$ we obtain [38, Theorem 2.2, 3.]

$$\overline{\text{conv}(\mathcal{R}_o)} = \overline{\mathcal{R}_o} = \bigcap_{\boldsymbol{\lambda} \in \mathbb{R}^3} \{\mathbf{x} \in \mathbb{R}^3 : \mathbf{x} \cdot \boldsymbol{\lambda} \leq \tilde{\psi}(\boldsymbol{\lambda})\}. \quad (140)$$

From the definition of \mathcal{R}_o , we clearly have $\tilde{\psi}((\lambda_1, \lambda_2, \lambda_3)) = \infty$ whenever $\lambda_1 < 0$, $\lambda_2 > 0$, or $\lambda_3 > 0$, and $\tilde{\psi}(\boldsymbol{\lambda}) = \psi(\boldsymbol{\lambda}) \triangleq \sup_{\mathbf{p} \in \mathcal{Q}} \boldsymbol{\lambda} \cdot \mathbf{F}(\mathbf{p})$ everywhere else. Thus,

$$\overline{\mathcal{R}_o} = \bigcap_{\boldsymbol{\lambda} \in \Lambda} \{\mathbf{x} \in \mathbb{R}^3 : \mathbf{x} \cdot \boldsymbol{\lambda} \leq \psi(\boldsymbol{\lambda})\}, \quad (141)$$

where Λ is the quadrant defined by $\lambda_1 \geq 0$, $\lambda_2 \leq 0$, and $\lambda_3 \leq 0$. We next show that for any $\boldsymbol{\lambda} \in \Lambda$,

$$\psi(\boldsymbol{\lambda}) = \max_{\mathbf{p} \in \mathcal{Q}(|\mathcal{X}|, |\mathcal{Z}|)} \boldsymbol{\lambda} \cdot \mathbf{F}(\mathbf{p}). \quad (142)$$

Choose arbitrary $\boldsymbol{\lambda} \in \Lambda$ and $\delta > 0$. We can find random variables $(\tilde{\mathbf{U}}, \tilde{\mathbf{X}}, \tilde{\mathbf{Z}}, \tilde{\mathbf{V}}) \sim \tilde{\mathbf{p}} \in \mathcal{Q}$ with $\boldsymbol{\lambda} \cdot \mathbf{F}(\tilde{\mathbf{p}}) \geq \psi(\boldsymbol{\lambda}) - \delta$. By compactness of $\mathcal{Q}(a, b)$ and continuity of \mathbf{F} , there are random variables $(\mathbf{U}, \mathbf{X}, \mathbf{Z}, \mathbf{V}) \sim \mathbf{p} \in \mathcal{Q}(|\mathcal{U}|, |\mathcal{V}|)$ with

$$\boldsymbol{\lambda} \cdot \mathbf{F}(\mathbf{p}) = \max_{\mathbf{p}' \in \mathcal{Q}(|\mathcal{U}|, |\mathcal{V}|)} \boldsymbol{\lambda} \cdot \mathbf{F}(\mathbf{p}') \geq \boldsymbol{\lambda} \cdot \mathbf{F}(\tilde{\mathbf{p}}) \geq \psi(\boldsymbol{\lambda}) - \delta. \quad (143)$$

We now show that there exists $\mathbf{p}'' \in \mathcal{Q}(|\mathcal{X}|, |\mathcal{Z}|)$ with

$$\boldsymbol{\lambda} \cdot \mathbf{F}(\mathbf{p}'') = \boldsymbol{\lambda} \cdot \mathbf{F}(\mathbf{p}). \quad (144)$$

By (143), $\boldsymbol{\lambda} \cdot \mathbf{F}(\mathbf{p}) = \max_{\mathbf{p}' \in \mathcal{Q}(|\mathcal{U}|, |\mathcal{V}|)} \boldsymbol{\lambda} \cdot \mathbf{F}(\mathbf{p}') = 0$ for any $\boldsymbol{\lambda} \in \Lambda$ with $\lambda_1 + \min(\lambda_2, \lambda_3) \leq 0$ as a consequence of the inequalities $F_1 \leq F_2$ and $F_1 \leq F_3$. Thus, we only need to show (144) for $\boldsymbol{\lambda} \in \Lambda$ with $\lambda_1 + \min(\lambda_2, \lambda_3) > 0$. To this end we use the perturbation method [39], [40] and perturb \mathbf{p} , obtaining

$$\mathbf{U}', \mathbf{X}', \mathbf{Z}', \mathbf{V}' \sim \mathbf{p}'(u, x, z, v) = \mathbf{p}(u, x, z, v)(1 + \varepsilon\phi(u)). \quad (145)$$

We require

$$1 + \varepsilon\phi(u) \geq 0, \quad (146)$$

$$\mathbb{E}[\phi(\mathbf{U})] = 0, \text{ and} \quad (147)$$

$$\mathbb{E}[\phi(\mathbf{U})|\mathbf{X} = x, \mathbf{Z} = z] = 0. \quad (148)$$

The conditions (146) and (147) ensure that \mathbf{p}' is a valid pmf and (148) implies $\mathbf{p}' \in \mathcal{Q}$. Observe that for any ϕ , there is an $\varepsilon_0 > 0$ such that (146) is satisfied for $\varepsilon \in [-\varepsilon_0, \varepsilon_0]$. Furthermore, (148) is equivalent to

$$\mathbb{E}[\phi(\mathbf{U})|\mathbf{X} = x] = 0 \quad (149)$$

because of the Markov chain $\mathbf{U} \ominus \mathbf{X} \ominus \mathbf{Z}$. If $|\mathcal{U}| \geq |\mathcal{X}| + 1$ there is a non-trivial solution to (147) and (149), which means there exists $\phi \not\equiv 0$ such that (146) to (148) are satisfied. We have

$$\begin{aligned} \boldsymbol{\lambda} \cdot \mathbf{F}(\mathbf{p}') &= \lambda_1[\mathbf{I}(\mathbf{X}; \mathbf{U}) - \mathbf{I}(\mathbf{UV}; \mathbf{XZ}) + \mathbf{H}(\mathbf{Z}) + \varepsilon\mathbf{H}(\mathbf{U})_\phi - \varepsilon\mathbf{H}(\mathbf{UX})_\phi \\ &\quad - \varepsilon\mathbf{H}(\mathbf{UV})_\phi + \varepsilon\mathbf{H}(\mathbf{UXZV})_\phi + \mathbf{H}(\mathbf{V}') - \mathbf{H}(\mathbf{Z}'\mathbf{V}')] \\ &\quad + \lambda_2[\mathbf{I}(\mathbf{X}; \mathbf{U}) + \varepsilon\mathbf{H}(\mathbf{U})_\phi - \varepsilon\mathbf{H}(\mathbf{UX})_\phi] \\ &\quad + \lambda_3[\mathbf{H}(\mathbf{Z}) + \mathbf{H}(\mathbf{V}') - \mathbf{H}(\mathbf{Z}'\mathbf{V}')]. \end{aligned} \quad (150)$$

We used the shorthand $\mathbf{H}(\mathbf{UX})_\phi \triangleq -\sum_{u,x} \mathbf{p}(u, x)\phi(u) \log \mathbf{p}(u, x)$ and analogous for other combinations of random

variables. By (143), we have $\frac{\partial^2}{\partial \varepsilon^2} \boldsymbol{\lambda} \cdot \mathbf{F}(\mathbf{p}')|_{\varepsilon=0} \leq 0$. Observe that

$$\frac{\partial}{\partial \varepsilon} (\mathbf{H}(\mathbf{V}') - \mathbf{H}(\mathbf{Z}'\mathbf{V}')) = \frac{\partial}{\partial \varepsilon} \sum_{z,v} p'(z,v) \log \frac{p'(z,v)}{p'(v)} \quad (151)$$

$$= \sum_{z,v} \frac{\partial p'(z,v)}{\partial \varepsilon} \log \frac{p'(z,v)}{p'(v)} + p'(z,v) \frac{p'(v)}{p'(z,v)} \frac{p'(v) \frac{\partial p'(z,v)}{\partial \varepsilon} - p'(z,v) \frac{\partial p'(v)}{\partial \varepsilon}}{p'(v)^2} \quad (152)$$

$$= \sum_{z,v} \frac{\partial p'(z,v)}{\partial \varepsilon} \log \frac{p'(z,v)}{p'(v)} + \frac{\partial p'(z,v)}{\partial \varepsilon} - \frac{p'(z,v) \frac{\partial p'(v)}{\partial \varepsilon}}{p'(v)} \quad (153)$$

and consequently

$$\frac{\partial^2}{\partial \varepsilon^2} \boldsymbol{\lambda} \cdot \mathbf{F}(\mathbf{p}') = (\lambda_1 + \lambda_3) \frac{\partial^2}{\partial \varepsilon^2} (\mathbf{H}(\mathbf{V}') - \mathbf{H}(\mathbf{Z}'\mathbf{V}')) \quad (154)$$

$$= (\lambda_1 + \lambda_3) \sum_{z,v} \frac{\partial p'(z,v)}{\partial \varepsilon} \frac{p'(v)}{p'(z,v)} \frac{p'(v) \frac{\partial p'(z,v)}{\partial \varepsilon} - p'(z,v) \frac{\partial p'(v)}{\partial \varepsilon}}{p'(v)^2} - \frac{\partial p'(v)}{\partial \varepsilon} \frac{p'(v) \frac{\partial p'(z,v)}{\partial \varepsilon} - p'(z,v) \frac{\partial p'(v)}{\partial \varepsilon}}{p'(v)^2} \quad (155)$$

$$= (\lambda_1 + \lambda_3) \sum_{z,v} \left(\frac{\partial p'(z,v)}{\partial \varepsilon} \right)^2 \frac{1}{p'(z,v)} - 2 \frac{\partial p'(z,v)}{\partial \varepsilon} \frac{\partial p'(v)}{\partial \varepsilon} \frac{1}{p'(v)} + \left(\frac{\partial p'(v)}{\partial \varepsilon} \right)^2 \frac{p'(z,v)}{p'(v)^2}. \quad (156)$$

Here we already used that $\frac{\partial^2 p'(v)}{\partial \varepsilon^2} \equiv \frac{\partial^2 p'(z,v)}{\partial \varepsilon^2} \equiv 0$. It is straightforward to calculate

$$\frac{\partial p'(v)}{\partial \varepsilon} = p_{\mathbf{V}}(v) \mathbb{E}[\phi(\mathbf{U}) | \mathbf{V} = v] \quad (157)$$

$$\frac{\partial p'(z,v)}{\partial \varepsilon} = p_{\mathbf{Z},\mathbf{V}}(z,v) \mathbb{E}[\phi(\mathbf{U}) | \mathbf{V} = v, \mathbf{Z} = z] \quad (158)$$

$$p'(z,v)|_{\varepsilon=0} = p_{\mathbf{Z},\mathbf{V}}(z,v) \quad (159)$$

$$p'(v)|_{\varepsilon=0} = p_{\mathbf{V}}(v) \quad (160)$$

and, thus, taking into account that $\lambda_1 + \lambda_3 > 0$,

$$0 \geq \sum_{z,v} p(z,v) (\mathbb{E}[\phi(\mathbf{U}) | \mathbf{V} = v, \mathbf{Z} = z] - \mathbb{E}[\phi(\mathbf{U}) | \mathbf{V} = v])^2, \quad (161)$$

which implies for any (z,v) with $p(z,v) > 0$,

$$\sum_u p_{\mathbf{U}|\mathbf{Z},\mathbf{V}}(u|z,v) \phi(u) = \sum_u p_{\mathbf{U}|\mathbf{V}}(u|v) \phi(u). \quad (162)$$

From (162) we can conclude

$$\mathbf{H}(\mathbf{V}') - \mathbf{H}(\mathbf{Z}'\mathbf{V}') = \sum_{z,v} p'(z,v) \log \frac{p'(z,v)}{p'(v)} \quad (163)$$

$$= \sum_{z,v,u} p(u,z,v) (1 + \varepsilon \phi(u)) \log \frac{\sum_{u'} p(u',z,v) (1 + \varepsilon \phi(u'))}{\sum_{u'} p(u',v) (1 + \varepsilon \phi(u'))} \quad (164)$$

$$= \sum_{z,v,u} p(u,z,v) (1 + \varepsilon \phi(u)) \log \frac{p(z,v) \sum_{u'} p(u'|z,v) (1 + \varepsilon \phi(u'))}{p(v) \sum_{u'} p(u'|v) (1 + \varepsilon \phi(u'))} \quad (165)$$

$$= \sum_{z,v,u} p(u, z, v)(1 + \varepsilon\phi(u)) \log \frac{p(z, v)(1 + \varepsilon \sum_{u'} p(u'|z, v)\phi(u'))}{p(v)(1 + \varepsilon \sum_{u'} p(u'|v)\phi(u'))} \quad (166)$$

$$\stackrel{(a)}{=} \sum_{z,v,u} p(u, z, v)(1 + \varepsilon\phi(u)) \log \frac{p(z, v)(1 + \varepsilon \sum_{u'} p(u'|v)\phi(u'))}{p(v)(1 + \varepsilon \sum_{u'} p(u'|v)\phi(u'))} \quad (167)$$

$$= \sum_{z,v,u} p(u, z, v)(1 + \varepsilon\phi(u)) \log \frac{p(z, v)}{p(v)} \quad (168)$$

$$= \sum_{z,v} p(z, v) \log \frac{p(z, v)}{p(v)} + \varepsilon \sum_{z,v,u} \phi(u) p(u, z, v) \log \frac{p(z, v)}{p(v)}, \quad (169)$$

where (a) follows from (162). Thus,

$$H(V') - H(Z'V') = H(V) - H(ZV) + \varepsilon H(V)_\phi - \varepsilon H(ZV)_\phi, \quad (170)$$

where we used

$$H(V)_\phi \triangleq - \sum_{u,v} p(u, v) \phi(u) \log p(v) \quad (171)$$

$$H(ZV)_\phi \triangleq - \sum_{u,z,v} p(u, z, v) \phi(u) \log p(z, v). \quad (172)$$

Substituting in (150) shows that $\lambda \cdot \mathbf{F}(p')$ is linear in ε . And by the optimality of p it must therefore be constant. We may now choose ε maximal, i.e., such that there is at least one $u \in \mathcal{U}$ with $p(u)(1 + \varepsilon\phi(u)) = 0$. This effectively reduces the cardinality of \mathcal{U}' by one and may be repeated until $\phi \equiv 0$, i.e. $|\mathcal{U}'| = |\mathcal{X}|$. The same process can be carried out for V and yields $p'' \in \mathcal{Q}(|\mathcal{X}|, |\mathcal{Z}|)$, such that (144) holds.

In the proof of \mathcal{R}'_o , we apply the support lemma [10, Appendix C] with $|\mathcal{X}| - 1$ test functions $f_x(p_{X'}) \triangleq p_{X'}(x)$ ($x \in \mathcal{X}$) and with the function

$$f(p_{X'}) \triangleq \lambda_1 \min(I(V; X), H(Z) - H(Z')) + \lambda_2(H(X) - H(X')) + \lambda_3 I(Z; V), \quad (173)$$

where $(Z', X') \sim p_{X'} p_{Z|X}$. Consequently there exists a random variable U' with $(U', X, Z, V) \sim p' \in \mathcal{Q}(|\mathcal{X}|, |\tilde{\mathcal{V}}|)$ and $\lambda \cdot \mathbf{F}(p') = \lambda \cdot \mathbf{F}(p)$. By applying the same argument to V , we obtain $p'' \in \mathcal{Q}(|\mathcal{X}|, |\mathcal{Z}|)$ such that (144) holds.

Due to (143) and (144) we now have

$$\lambda \cdot \mathbf{F}(p'') = \lambda \cdot \mathbf{F}(p) \geq \psi(\lambda) - \delta. \quad (174)$$

Since this holds for arbitrary $\delta > 0$ and since $\mathcal{Q}(|\mathcal{X}|, |\mathcal{Z}|)$ is compact, (142) holds. Now (141) implies

$$\overline{\mathcal{R}_o} = \overline{\text{conv}(\mathbf{F}(\mathcal{Q}(|\mathcal{X}|, |\mathcal{Z}|)) + \mathcal{O})} \quad (175)$$

$$\stackrel{(a)}{=} \overline{\text{conv}(\mathbf{F}(\mathcal{Q}(|\mathcal{X}|, |\mathcal{Z}|)) + \mathcal{O})} \quad (176)$$

$$\stackrel{(b)}{=} \overline{\text{conv}(\mathbf{F}(\mathcal{Q}(|\mathcal{X}|, |\mathcal{Z}|)) + \mathcal{O})} \quad (177)$$

$$\stackrel{(c)}{\subseteq} \mathbf{F}(\mathcal{Q}) + \mathcal{O} \quad (178)$$

$$= \mathcal{R}_o, \quad (179)$$

where (a) follows from [41, Theorem 1.1.2] and (b) is a consequence of [42, Exercise 1.3(e), Theorem 1.13(b)], considering that $\mathbf{F}(\mathcal{Q}(|\mathcal{X}|, |\mathcal{Z}|))$ and therefore also its convex hull is compact [38, Theorem 2.3, 4.]. The relation (c) is a consequence of $\mathcal{Q}(|\mathcal{X}|, |\mathcal{Z}|) \subseteq \mathcal{Q}$ and the convexity of $\mathbf{F}(\mathcal{Q})$.

D. Proof of Lemma 9

Fix $\varepsilon' > 0$ and $n \in \mathbb{N}$. For n sufficiently large we find $M_1, M_2 \in \mathbb{N}$ satisfying (26) and (27). We can thus apply Lemma 26. Denote the two codebooks $\mathcal{C}_U \triangleq (\mathbf{U}_i)_{i=\{1,2,\dots,M_1\}}$ and $\mathcal{C}_V \triangleq (\mathbf{V}_j)_{j=\{1,2,\dots,M_2\}}$, which are drawn independently uniform from $\mathcal{T}_{[U]\delta}^n$ and from $\mathcal{T}_{[V]\delta}^n$, respectively, where $\delta > 0$ is sufficiently small. Denoting the

resulting randomized decoded values as \mathbf{U}^* and \mathbf{V}^* , we have for n large enough that

$$P_e \triangleq \mathbb{P}\left\{(\mathbf{U}^*, \mathbf{X}, \mathbf{Z}, \mathbf{V}^*) \notin \mathcal{T}_{[\mathbf{U}\mathbf{X}\mathbf{Z}\mathbf{V}]\delta}^n\right\} \leq \varepsilon'. \quad (180)$$

We next analyze the random quantity $L \triangleq \sum_{i,j=1}^{M_1, M_2} \mathbb{1}_{\mathcal{T}_{[\mathbf{U}\mathbf{V}]\delta}^n}(\hat{\mathbf{U}}_i, \hat{\mathbf{V}}_j)$. For n large enough,

$$\mathbb{E}[L] = \sum_{i,j=1}^{M_1, M_2} \mathbb{E}\left[\mathbb{1}_{\mathcal{T}_{[\mathbf{U}\mathbf{V}]\delta}^n}(\hat{\mathbf{U}}_i, \hat{\mathbf{V}}_j)\right] \quad (181)$$

$$= \sum_{i,j=1}^{M_1, M_2} \frac{|\mathcal{T}_{[\mathbf{U}\mathbf{V}]\delta}^n|}{|\mathcal{T}_{[\mathbf{U}]\delta}^n| |\mathcal{T}_{[\mathbf{V}]\delta}^n|} \quad (182)$$

$$\stackrel{(a)}{\leq} M_1 M_2 \frac{e^{n(H(\mathbf{U}\mathbf{V}) + \varepsilon_1(\delta))}}{e^{n(H(\mathbf{U}) + H(\mathbf{V}) - \varepsilon_2(\delta))}} \quad (183)$$

$$\leq M_1 M_2 e^{-n(I(\mathbf{U}; \mathbf{V}) - \varepsilon(\delta))} \quad (184)$$

where $\varepsilon(\delta) = \varepsilon_1(\delta) + \varepsilon_2(\delta)$ goes to zero as $\delta \rightarrow 0$. Here (a) follows from parts 1 and 2 of Lemma 25. If δ is sufficiently small, we can choose $\varepsilon_1, \varepsilon_2 < \varepsilon$ such that $\varepsilon_1 + \varepsilon_2 + \varepsilon_3(\delta) < \varepsilon$. Requiring $M_1 \leq e^{n(I(\mathbf{U}; \mathbf{X}) + \varepsilon_1)}$ and $M_2 \leq e^{n(I(\mathbf{V}; \mathbf{Z}) + \varepsilon_2)}$, we have from (184)

$$\mathbb{E}[L] \leq \exp(n(I(\mathbf{U}\mathbf{V}; \mathbf{X}\mathbf{Z}) + \varepsilon_1 + \varepsilon_2 + \varepsilon_3(\delta))) \quad (185)$$

and we know from Markov's inequality that for n large enough

$$\mathbb{P}\{L \geq \exp(n(I(\mathbf{U}\mathbf{V}; \mathbf{X}\mathbf{Z}) + \varepsilon))\} \leq \exp(n(\varepsilon_1 + \varepsilon_2 + \varepsilon_3(\delta) - \varepsilon)) \leq \varepsilon'. \quad (186)$$

Define the error events $\mathcal{E}_1 = \{(\mathbf{U}^*, \mathbf{X}, \mathbf{Z}, \mathbf{V}^*) \notin \mathcal{T}_{[\mathbf{U}\mathbf{X}\mathbf{Z}\mathbf{V}]\delta}^n\}$ and $\mathcal{E}_2 = \{L > \exp(n(I(\mathbf{U}\mathbf{V}; \mathbf{X}\mathbf{Z}) + \varepsilon))\}$. Markov's inequality implies

$$\mathbb{P}\{\mathbb{P}\{\mathcal{E}_1 | \mathcal{C}_U, \mathcal{C}_V\} \geq \sqrt{\varepsilon'}\} \leq \sqrt{\varepsilon'}, \quad (187)$$

$$\mathbb{P}\{\mathbb{P}\{\mathcal{E}_2 | \mathcal{C}_U, \mathcal{C}_V\} \geq \sqrt{\varepsilon'}\} \leq \sqrt{\varepsilon'}. \quad (188)$$

With $\varepsilon' = \min(\varepsilon^2, \frac{1}{8})$, the union bound guarantees that with probability at least $1 - 2\sqrt{\varepsilon'}$ our random coding scheme yields a code $\mathcal{C}_u = (\mathbf{u}_i)_{i=\{1,2,\dots,M_1\}}$, $\mathcal{C}_v = (\mathbf{v}_j)_{j=\{1,2,\dots,M_2\}}$ and deterministic encoding functions $f: \mathcal{X}^n \rightarrow \mathcal{C}_u$, $g: \mathcal{Z}^n \rightarrow \mathcal{C}_v$, such that

$$\mathbb{P}\{\mathcal{E}_1 | \mathcal{C}_V = \mathcal{C}_v, \mathcal{C}_U = \mathcal{C}_u\} = \mathbb{P}\{(f(\mathbf{X}), \mathbf{X}, \mathbf{Z}, g(\mathbf{Z})) \notin \mathcal{T}_{[\mathbf{U}\mathbf{X}\mathbf{Z}\mathbf{V}]\delta}^n\} \leq \varepsilon, \quad (189)$$

$$\sum_{i,j=1}^{M_1, M_2} \mathbb{1}_{\mathcal{T}_{[\mathbf{U}\mathbf{V}]\delta}^n}(\mathbf{u}_i, \mathbf{v}_j) \leq \exp(n(I(\mathbf{U}\mathbf{V}; \mathbf{X}\mathbf{Z}) + \varepsilon)). \quad (190)$$

Pick any such code and define $\mathcal{C}_i = f^{-1}(\{\mathbf{u}_i\}) \cap \mathcal{T}_{[\mathbf{X}|\mathbf{U}]\delta}^n(\mathbf{u}_i)$ if $\mathbf{u}_i \neq \mathbf{u}_{i'}$ for all $i' < i$ and $\mathcal{C}_i = \emptyset$ otherwise. \mathcal{D}_j is defined accordingly. The conditions (28) and (29) now follow directly from (189) and (190).

E. Proof of Proposition 6

We only need to show $\text{conv}(\mathcal{S}_i) = \text{conv}(\mathcal{R}_i)$ as the cardinality bound $|\mathcal{Q}| \leq 3$ follows directly from the strengthened Carathéodory theorem [43, Theorem 18(ii)] because $\text{conv}(\mathcal{R}_i)$ is the convex hull of a connected set in \mathbb{R}^3 . We will only show the cardinality bound $|\mathcal{U}| \leq |\mathcal{X}|$ as the bound for $|\mathcal{V}|$ follows analogously. We note that the weaker bounds $|\mathcal{U}| \leq |\mathcal{X}| + 1$ and $|\mathcal{V}| \leq |\mathcal{Z}| + 1$ can be shown directly using the convex cover method [10, Appendix C], [19], [44]. Define the continuous vector-valued function

$$\mathbf{F}(\mathbf{p}_{\tilde{\mathbf{U}}, \tilde{\mathbf{X}}, \tilde{\mathbf{Z}}, \tilde{\mathbf{V}}}) \triangleq (I(\tilde{\mathbf{U}}; \tilde{\mathbf{V}}), I(\tilde{\mathbf{X}}; \tilde{\mathbf{U}}), I(\tilde{\mathbf{Z}}; \tilde{\mathbf{V}})). \quad (191)$$

Define the compact, connected sets of pmfs

$$\mathcal{Q} \triangleq \{p_{\tilde{U}, \tilde{X}, \tilde{Z}, \tilde{V}} : p_{\tilde{U}, \tilde{X}, \tilde{Z}, \tilde{V}} = p_{\tilde{U}|\tilde{X}} p_{\tilde{X}, \tilde{Z}|\tilde{V}}, \tilde{\mathcal{U}} = \{0, \dots, |\mathcal{X}|\}, \tilde{\mathcal{V}} = \{0, \dots, |\mathcal{Z}|\}\}, \quad (192)$$

$$\mathcal{Q}' \triangleq \{p_{\tilde{U}, \tilde{X}, \tilde{Z}, \tilde{V}} \in \mathcal{Q} : \tilde{\mathcal{U}} = \{1, \dots, |\mathcal{X}|\}\}. \quad (193)$$

To complete the proof of the proposition, it suffices to show

$$\text{conv}(\mathbf{F}(\mathcal{Q})) \subseteq \text{conv}(\mathbf{F}(\mathcal{Q}')) \quad (194)$$

since we then have

$$\text{conv}(\mathcal{R}_i) = \text{conv}(\mathbf{F}(\mathcal{Q}) + \mathcal{O}) \quad (195)$$

$$\stackrel{(a)}{\subseteq} \text{conv}(\mathbf{F}(\mathcal{Q})) + \text{conv}(\mathcal{O}) \quad (196)$$

$$= \text{conv}(\mathbf{F}(\mathcal{Q})) + \mathcal{O} \quad (197)$$

$$\stackrel{(b)}{\subseteq} \text{conv}(\mathbf{F}(\mathcal{Q}')) + \mathcal{O} \quad (198)$$

$$= \text{conv}(\mathbf{F}(\mathcal{Q}') + \mathcal{O}) \quad (199)$$

$$= \text{conv}(\mathcal{S}_i), \quad (200)$$

where (a) follows from [41, Theorem 1.1.2], (b) from (194), and we used $\mathcal{O} = (\mathbb{R}_- \times \mathbb{R}_+ \times \mathbb{R}_+)$. The region $\mathbf{F}(\mathcal{Q}) \subseteq \mathbb{R}^3$ is compact and connected [45, Theorem 26.5], [46, Theorem 4.22]. Therefore, its convex hull $\text{conv}(\mathbf{F}(\mathcal{Q}))$ is compact [47, Corollary 5.33] and can be represented as an intersection of halfspaces in the following manner [38, Proposition 2.2, 3.]: For $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3) \in \mathbb{R}^3$, define $V(\boldsymbol{\lambda}) \triangleq \max_{\mathbf{x} \in \mathbf{F}(\mathcal{Q})} \boldsymbol{\lambda} \cdot \mathbf{x}$, then

$$\text{conv}(\mathbf{F}(\mathcal{Q})) = \bigcap_{\boldsymbol{\lambda} \in \mathbb{R}^3} \{\mathbf{x} \in \mathbb{R}^3 : \boldsymbol{\lambda} \cdot \mathbf{x} \leq V(\boldsymbol{\lambda})\}. \quad (201)$$

With the same reasoning we obtain

$$\text{conv}(\mathbf{F}(\mathcal{Q}')) = \bigcap_{\boldsymbol{\lambda} \in \mathbb{R}^3} \{\mathbf{x} \in \mathbb{R}^3 : \boldsymbol{\lambda} \cdot \mathbf{x} \leq V'(\boldsymbol{\lambda})\}, \quad (202)$$

where $V'(\boldsymbol{\lambda}) \triangleq \max_{\mathbf{x} \in \mathbf{F}(\mathcal{Q}')} \boldsymbol{\lambda} \cdot \mathbf{x}$. We next show $V'(\boldsymbol{\lambda}) \geq V(\boldsymbol{\lambda})$ which already implies (194) due to (201) and (202).

Let $\mathcal{X}' \triangleq \mathcal{X} \setminus \{x\}$ where $x \in \mathcal{X}$ is arbitrary. Define the test function $t_x(p_{\tilde{X}}) \triangleq p_{\tilde{X}}(x)$ for $x \in \mathcal{X}'$ and abbreviate $\mathbf{t} = (t_x)_{x \in \mathcal{X}'}$. Choose any $\boldsymbol{\lambda} \in \mathbb{R}^3$ and fix (U, X, Z, V) that achieve $\boldsymbol{\lambda} \cdot \mathbf{F}(p_{U, X, Z, V}) = V(\boldsymbol{\lambda})$. Define the continuous function

$$f(p_{\tilde{X}}) \triangleq \lambda_1(H(X) - H(\tilde{X})) + \lambda_2 I(Z; V) + \lambda_3(H(V) - H(\tilde{V})) \quad (203)$$

where $(\tilde{V}, \tilde{Z}, \tilde{X}) \sim p_{V|Z} p_{Z|X} p_{\tilde{X}}$. Obviously $((p_X(x))_{x \in \mathcal{X}'}, V(\boldsymbol{\lambda}))$ lies in the convex hull of the compact, connected set $(\mathbf{t}, f)(\mathcal{Q}(\mathcal{X}'))$. Therefore, by the strengthened Carathéodory theorem [43, Theorem 18(ii)], $|\mathcal{X}'|$ points suffice, i.e., there exists a random variable U' with $|\mathcal{U}'| = |\mathcal{X}'|$ and thus $p_{U', X, Z, V} \in \mathcal{Q}'$, such that $\mathbb{E}[f(p_{X|U'}(\cdot | U'))] = \boldsymbol{\lambda} \cdot \mathbf{F}(p_{U', X, Z, V}) = V(\boldsymbol{\lambda})$. This shows $V'(\boldsymbol{\lambda}) \geq V(\boldsymbol{\lambda})$.

By applying the same reasoning to V , one can show that $|\mathcal{V}| = |\mathcal{Z}|$ is sufficient.

F. Proof of Theorem 13

If $(\mu_{\Omega}, R_{\mathcal{K}}) \in \mathcal{R}$ we obtain an $(n, R_{\mathcal{K}})$ code $f_{\mathcal{K}}$ for some $n \in \mathbb{N}$ such that (64) holds. Define $U_{\mathcal{A}} \triangleq f_{\mathcal{K}}(\mathbf{X}_k)_{k \in \mathcal{A}}$ and the auxiliary random variables $U_{k,l} \triangleq (f_{\mathcal{K}}(\mathbf{X}_k), \mathbf{X}_k^{l-1})$ for $k \in \mathcal{K}$ and $l \in \{1, 2, \dots, n\}$. For any $k \in \mathcal{K}$ we then have

$$nR_k \geq H(U_k) \quad (204)$$

$$= I(U_k; \mathbf{X}_k) \quad (205)$$

$$= \sum_{l=1}^n I(U_k; \mathbf{X}_{k,l} | \mathbf{X}_k^{l-1}) \quad (206)$$

$$= \sum_{l=1}^n I(U_k, \mathbf{X}_k^{l-1}; \mathbf{X}_{k,l}) \quad (207)$$

$$= \sum_{l=1}^n I(U_{k,l}; \mathbf{X}_{k,l}). \quad (208)$$

Furthermore, for any pair $(\mathcal{A}, \mathcal{B}) \in \Omega$

$$n\mu \leq I(U_{\mathcal{A}}; U_{\mathcal{B}}) \quad (209)$$

$$= I(U_{\mathcal{A}}; \mathbf{X}_{\mathcal{A}}) - I(U_{\mathcal{A}}; \mathbf{X}_{\mathcal{A}} | U_{\mathcal{B}}) \quad (210)$$

$$= I(U_{\mathcal{A}}; \mathbf{X}_{\mathcal{A}}) + I(U_{\mathcal{B}}; \mathbf{X}_{\mathcal{B}}) - I(U_{\mathcal{A}}; \mathbf{X}_{\mathcal{A}} | U_{\mathcal{B}}) - I(U_{\mathcal{B}}; \mathbf{X}_{\mathcal{B}}) \quad (211)$$

$$= I(U_{\mathcal{A}}; \mathbf{X}_{\mathcal{A}}) + I(U_{\mathcal{B}}; \mathbf{X}_{\mathcal{B}}) - I(U_{\mathcal{A}}; \mathbf{X}_{\mathcal{A}} \mathbf{X}_{\mathcal{B}} | U_{\mathcal{B}}) - I(U_{\mathcal{B}}; \mathbf{X}_{\mathcal{A}} \mathbf{X}_{\mathcal{B}}) \quad (212)$$

$$= I(U_{\mathcal{A}}; \mathbf{X}_{\mathcal{A}}) + I(U_{\mathcal{B}}; \mathbf{X}_{\mathcal{B}}) - I(U_{\mathcal{A}} U_{\mathcal{B}}; \mathbf{X}_{\mathcal{A}} \mathbf{X}_{\mathcal{B}}) \quad (213)$$

$$\leq \sum_{l=1}^n \left[I(U_{\mathcal{A},l}; \mathbf{X}_{\mathcal{A},l}) + I(U_{\mathcal{B},l}; \mathbf{X}_{\mathcal{B},l}) - I(U_l, V_l; \mathbf{X}_{\mathcal{A},l}, \mathbf{X}_{\mathcal{B},l}) \right]. \quad (214)$$

Now a standard time-sharing argument shows $\mathcal{R} \subseteq \mathcal{R}_o$. To see $\mathcal{R}_o \subseteq \mathcal{R}'_o$ note that

$$I(U_{\mathcal{A}}; \mathbf{X}_{\mathcal{A}}) + I(U_{\mathcal{B}}; \mathbf{X}_{\mathcal{B}}) - I(U_{\mathcal{A}} U_{\mathcal{B}}; \mathbf{X}_{\mathcal{A}} \mathbf{X}_{\mathcal{B}}) \quad (215)$$

$$= I(U_{\mathcal{B}}; \mathbf{X}_{\mathcal{B}}) - I(U_{\mathcal{B}}; \mathbf{X}_{\mathcal{A}} \mathbf{X}_{\mathcal{B}} | U_{\mathcal{A}}) \quad (216)$$

$$= I(U_{\mathcal{A}}; \mathbf{X}_{\mathcal{B}}) + I(U_{\mathcal{B}}; \mathbf{X}_{\mathcal{B}}) - I(U_{\mathcal{A}}; \mathbf{X}_{\mathcal{B}}) - I(U_{\mathcal{B}}; \mathbf{X}_{\mathcal{A}} \mathbf{X}_{\mathcal{B}} | U_{\mathcal{A}}) \quad (217)$$

$$= I(U_{\mathcal{A}}; \mathbf{X}_{\mathcal{B}}) + I(U_{\mathcal{B}}; \mathbf{X}_{\mathcal{B}}) - I(U_{\mathcal{A}}; \mathbf{X}_{\mathcal{B}}) - I(U_{\mathcal{B}}; \mathbf{X}_{\mathcal{B}} | U_{\mathcal{A}}) - I(U_{\mathcal{B}}; \mathbf{X}_{\mathcal{A}} | \mathbf{X}_{\mathcal{B}} U_{\mathcal{A}}) \quad (218)$$

$$= I(U_{\mathcal{A}}; \mathbf{X}_{\mathcal{B}}) + I(U_{\mathcal{B}}; \mathbf{X}_{\mathcal{B}}) - I(U_{\mathcal{A}} U_{\mathcal{B}}; \mathbf{X}_{\mathcal{B}}) - I(U_{\mathcal{B}}; \mathbf{X}_{\mathcal{A}} | \mathbf{X}_{\mathcal{B}} U_{\mathcal{A}}) \quad (219)$$

$$= I(U_{\mathcal{A}}; \mathbf{X}_{\mathcal{B}}) - I(U_{\mathcal{A}}; \mathbf{X}_{\mathcal{B}} | U_{\mathcal{B}}) - I(U_{\mathcal{B}}; \mathbf{X}_{\mathcal{A}} | \mathbf{X}_{\mathcal{B}} U_{\mathcal{A}}) \quad (220)$$

$$\leq I(U_{\mathcal{A}}; \mathbf{X}_{\mathcal{B}}). \quad (221)$$

G. Proof of Proposition 15

Pick an arbitrary $k \in \mathcal{K}$. For $(\mathcal{A}, \mathcal{B}) \in \Omega$ with $k \in \mathcal{B}$ we can write $H(\mathbf{X}_{\mathcal{A}} | U_{\mathcal{B}}) = \mathbb{E}[f_{\mathcal{A},\mathcal{B}}(\mathbf{p}_{\mathbf{X}_k | U_k}(\cdot | U_k))]$ where

$$f_{\mathcal{A},\mathcal{B}}(\mathbf{p}_{\mathbf{X}_k | U_k}(\cdot | u_k)) \triangleq H(\mathbf{X}_{\mathcal{A}} | U_{\mathcal{B} \setminus k}, U_k = u_k). \quad (222)$$

Furthermore, if $k \notin \mathcal{A}$, $H(U_{\mathcal{A}} | U_{\mathcal{B}}) = \mathbb{E}[g_{\mathcal{A},\mathcal{B}}(\mathbf{p}_{\mathbf{X}_k | U_k}(\cdot | U_k))]$ where

$$g_{\mathcal{A},\mathcal{B}}(\mathbf{p}_{\mathbf{X}_k | U_k}(\cdot | u_k)) \triangleq H(U_{\mathcal{A}} | U_{\mathcal{B} \setminus k}, U_k = u_k). \quad (223)$$

Observe that both $f_{\mathcal{A},\mathcal{B}}$ and $g_{\mathcal{A},\mathcal{B}}$ are continuous functions of $\mathbf{p}_{\mathbf{X}_k | U_k}(\cdot | u_k)$. Apply the support lemma [10, Appendix C] with the functions $f_{\mathcal{A},\mathcal{B}}$ and $g_{\mathcal{A},\mathcal{B}}$ for all pairs $(\mathcal{A}, \mathcal{B}) \in \Omega$ such that $k \in \mathcal{B}$, and $|\mathcal{X}_k| - 1$ test functions, which guarantee that the marginal distribution $\mathbf{p}_{\mathbf{X}_k}$ does not change. We obtain a new random variable U'_k with $H(\mathbf{X}_{\mathcal{A}} | U_{\mathcal{B} \setminus k} U'_k) = H(\mathbf{X}_{\mathcal{A}} | U_{\mathcal{B}})$ and $H(U_{\mathcal{A}} | U_{\mathcal{B} \setminus k} U'_k) = H(U_{\mathcal{A}} | U_{\mathcal{B}})$ if $k \notin \mathcal{A}$. By rewriting (66) to (68) in terms of conditional entropies, it is evident that the defining inequalities for \mathcal{R}_1 remain the same when replacing U_k by U'_k . U'_k satisfies the required cardinality bound²

$$|\mathcal{U}'_k| \leq |\mathcal{X}_k| - 1 + 2(2^K - 1)2^{K-1} \quad (224)$$

$$= |\mathcal{X}_k| - 1 + 2^{2K} - 2^K \quad (225)$$

$$\leq |\mathcal{X}_k| + 4^K. \quad (226)$$

The same process is repeated for every $k \in \mathcal{K}$.

²There are $(2^K - 1)$ ways to choose \mathcal{A} and 2^{K-1} ways to choose \mathcal{B} .

H. Proof of Lemma 18

We can interpret a reproduction $p_{\hat{\mathbf{Y}}_{\mathcal{B}}|\mathbf{U}_{\mathcal{A}}}(\cdot|u_{\mathcal{A}}) \triangleq g_{\mathcal{A},\mathcal{B}}(u_{\mathcal{A}})$ as the conditional probability distribution of $\hat{\mathbf{Y}}_{\mathcal{B}}$ given $\mathbf{U}_{\mathcal{A}} = u_{\mathcal{A}}$. Note that this is in general not a product distribution. We calculate

$$\mathbb{E}[\zeta_{\mathcal{B}}(\mathbf{Y}_{\mathcal{B}}, g_{\mathcal{B}}(\mathbf{U}_{\mathcal{A}})) | \mathbf{U}_{\mathcal{A}} = u_{\mathcal{A}}] = H(\mathbf{Y}_{\mathcal{B}}) + \frac{1}{n} \sum_{\mathbf{y}_{\mathcal{B}} \in \mathcal{Y}_{\mathcal{B}}^n} p_{\mathbf{Y}_{\mathcal{B}}|\mathbf{U}_{\mathcal{A}}}(\mathbf{y}_{\mathcal{B}}|u_{\mathcal{A}}) \log p_{\hat{\mathbf{Y}}_{\mathcal{B}}|\mathbf{U}_{\mathcal{A}}}(\mathbf{y}_{\mathcal{B}}|u_{\mathcal{A}}) \quad (227)$$

$$= H(\mathbf{Y}_{\mathcal{B}}) - \frac{1}{n} \sum_{\mathbf{y}_{\mathcal{B}} \in \mathcal{Y}_{\mathcal{B}}^n} p_{\mathbf{Y}_{\mathcal{B}}|\mathbf{U}_{\mathcal{A}}}(\mathbf{y}_{\mathcal{B}}|u_{\mathcal{A}}) \log \frac{p_{\mathbf{Y}_{\mathcal{B}}|\mathbf{U}_{\mathcal{A}}}(\mathbf{y}_{\mathcal{B}}|u_{\mathcal{A}})}{p_{\hat{\mathbf{Y}}_{\mathcal{B}}|\mathbf{U}_{\mathcal{A}}}(\mathbf{y}_{\mathcal{B}}|u_{\mathcal{A}})} - \frac{1}{n} H(\mathbf{Y}_{\mathcal{B}} | \mathbf{U}_{\mathcal{A}} = u_{\mathcal{A}}) \quad (228)$$

$$= H(\mathbf{Y}_{\mathcal{B}}) - \frac{1}{n} D_{\text{KL}}(p_{\mathbf{Y}_{\mathcal{B}}|\mathbf{U}_{\mathcal{A}}}(\cdot|u_{\mathcal{A}}) \| p_{\hat{\mathbf{Y}}_{\mathcal{B}}|\mathbf{U}_{\mathcal{A}}}(\cdot|u_{\mathcal{A}})) - \frac{1}{n} H(\mathbf{Y}_{\mathcal{B}} | \mathbf{U}_{\mathcal{A}} = u_{\mathcal{A}}) \quad (229)$$

$$\leq \frac{1}{n} \left[H(\mathbf{Y}_{\mathcal{B}}) - H(\mathbf{Y}_{\mathcal{B}} | \mathbf{U}_{\mathcal{A}} = u_{\mathcal{A}}) \right], \quad (230)$$

where in the last step we used the non-negativity of Kullback-Leibler divergence. Equality in (230) is obtained with $g_{\mathcal{B}}(u_{\mathcal{A}}) = p_{\mathbf{Y}_{\mathcal{B}}|\mathbf{U}_{\mathcal{A}}}(\cdot|u_{\mathcal{A}})$. By averaging over $u_{\mathcal{A}}$ we obtain the desired result.

I. Proof of Lemma 20

Fix $0 < \varepsilon', \varepsilon'' < \varepsilon$ and $\tilde{R}_{\mathcal{K}} \in \mathbb{R}_+^K$ as $\tilde{R}_k = I(\mathbf{X}_k; \mathbf{U}_k) + \varepsilon''/2$ for each $k \in \mathcal{K}$.

- **Encoding:** For $n \in \mathbb{N}$ define $\tilde{M}_k \triangleq e^{n\tilde{R}_k}$ and $\tilde{\mathcal{M}}_k \triangleq \{1, 2, \dots, \tilde{M}_k\}$. We apply Lemma 26 and consider the random codebooks $\mathcal{C}_k \triangleq (\mathbf{V}_i^{(k)})_{i \in \tilde{\mathcal{M}}_k}$, which are drawn independently uniform from $\mathcal{T}_{[\mathbf{U}_k]_{\delta}}^n$ for each $k \in \mathcal{K}$. Denote the resulting randomized coding functions as $\tilde{W}_k = \tilde{f}_k(\mathbf{X}_k, \mathcal{C}_k)$ and the corresponding decoded value as $\tilde{\mathbf{U}}_k \triangleq \mathbf{V}_{\tilde{W}_k}^{(k)}$. If n is chosen large enough and δ small enough we have therefore

$$P_e \triangleq \mathbb{P}\{(\tilde{\mathbf{U}}_{\mathcal{K}}, \mathbf{X}_{\mathcal{K}}) \notin \mathcal{T}_{[\mathbf{U}_{\mathcal{K}}\mathbf{X}_{\mathcal{K}}]_{\delta}}^n\} \leq \varepsilon'. \quad (231)$$

Next, we introduce (deterministic) binning. If $R_k < I(\mathbf{X}_k; \mathbf{U}_k)$, partition $\tilde{\mathcal{M}}_k$ into $M_k \triangleq e^{n(R_k + \varepsilon'')}$ equally sized, consecutive bins, each of size $e^{n\Delta_k}$ with

$$\Delta_k \triangleq \tilde{R}_k - R_k - \varepsilon'' = I(\mathbf{X}_k; \mathbf{U}_k) - R_k - \frac{\varepsilon''}{2}. \quad (232)$$

The deterministic function $\beta_k: \tilde{\mathcal{M}}_k \rightarrow \mathcal{M}_k$ maps a codeword to the index of the bin in $\mathcal{M}_k \triangleq \{1, 2, \dots, M_k\}$ to which it belongs. Now use the randomized encoding function $f_k \triangleq \beta_k \circ \tilde{f}_k$. If $R_k \geq I(\mathbf{X}_k; \mathbf{U}_k)$, we do not require binning and let β_k be the identity on $\tilde{\mathcal{M}}_k$ and hence $f_k \triangleq \tilde{f}_k$.

- **Decoding:** Given the codebooks, the decoding procedure $g_{\mathcal{A}_a, \mathcal{A}_b}: \mathcal{M}_{\mathcal{A}_a} \rightarrow \mathcal{U}_{\mathcal{A}_b}^n$ for each $\emptyset \neq \mathcal{A}_b \subseteq \mathcal{A}_a \subseteq \mathcal{K}$ is carried out as follows: Given $m_{\mathcal{A}_a} \in \mathcal{M}_{\mathcal{A}_a}$, let $\tilde{m}_{\mathcal{A}_a} \triangleq \beta_{\mathcal{A}_a}^{-1}(m_{\mathcal{A}_a}) \subseteq \tilde{\mathcal{M}}_{\mathcal{A}_a}$ be all indices that are in the bins $m_{\mathcal{A}_a}$. Consider only the typical sequences $\mathbf{V}_{\tilde{m}_{\mathcal{A}_a}}^{(\mathcal{A}_a)} \cap \mathcal{T}_{[\mathbf{U}_{\mathcal{A}_a}]_{\delta}}^n \triangleq \Phi \subseteq \mathcal{U}_{\mathcal{A}_a}^n$. Denote the restriction of Φ to the coordinates \mathcal{A}_b as $[\Phi]_{\mathcal{A}_b}$. If $\Phi \neq \emptyset$, choose the lexicographically smallest element of $[\Phi]_{\mathcal{A}_b}$, otherwise choose the lexicographically smallest element of $[\mathbf{V}_{\tilde{m}_{\mathcal{A}_a}}^{(\mathcal{A}_a)}]_{\mathcal{A}_b}$.

Let $\mathcal{A}_a, \mathcal{A}_b, \mathcal{B}_a, \mathcal{B}_b \subseteq \mathcal{K}$ be sets of indices, such that the conditions of part 1 are satisfied. Using $W_k \triangleq f_k(\mathbf{X}_k, \mathcal{C}_k)$ and the randomized decodings $\hat{\mathbf{U}}_1 \triangleq g_{\mathcal{A}_a, \mathcal{A}_b}(W_{\mathcal{A}_a}, \mathcal{C}_{\mathcal{A}_a})$ and $\hat{\mathbf{U}}_2 \triangleq g_{\mathcal{B}_a, \mathcal{B}_b}(W_{\mathcal{B}_a}, \mathcal{C}_{\mathcal{B}_a})$, consider the error event $\mathcal{E}_0 \triangleq \{(\hat{\mathbf{U}}_1, \mathbf{X}_{\mathcal{A}_a}, \mathbf{X}_{\mathcal{B}_a}, \hat{\mathbf{U}}_2) \notin \mathcal{T}_{[\mathbf{U}_{\mathcal{A}_b}\mathbf{X}_{\mathcal{A}_a}\mathbf{X}_{\mathcal{B}_a}\mathbf{U}_{\mathcal{B}_b}]_{\delta}}^n\}$. Defining the other events

$$\mathcal{E}_1 \triangleq \{(\tilde{\mathbf{U}}_{\mathcal{A}_a}, \mathbf{X}_{\mathcal{A}_a}, \mathbf{X}_{\mathcal{B}_a}, \tilde{\mathbf{U}}_{\mathcal{B}_a}) \notin \mathcal{T}_{[\mathbf{U}_{\mathcal{A}_b}\mathbf{X}_{\mathcal{A}_a}\mathbf{X}_{\mathcal{B}_a}\mathbf{U}_{\mathcal{B}_b}]_{\delta}}^n\}, \quad (233)$$

$$\mathcal{E}_2 \triangleq \left\{ \left| \left[\mathbf{V}_{\tilde{\mathcal{M}}_{\mathcal{A}_a}}^{(\mathcal{A}_a)} \cap \mathcal{T}_{[\mathbf{U}_{\mathcal{A}_a}]_{\delta}}^n \right]_{\mathcal{A}_b} \right| > 1 \right\}, \quad (234)$$

$$\mathcal{E}_3 \triangleq \left\{ \left| \left[\mathbf{V}_{\tilde{\mathcal{M}}_{\mathcal{B}_a}}^{(\mathcal{B}_a)} \cap \mathcal{T}_{[\mathbf{U}_{\mathcal{B}_a}]_{\delta}}^n \right]_{\mathcal{B}_b} \right| > 1 \right\}, \quad (235)$$

where we defined the random set of indices $\tilde{\mathcal{M}}_{\mathcal{A}} \triangleq \beta_{\mathcal{A}}^{-1}(W_{\mathcal{A}})$. We clearly have $\mathcal{E}_0 \subseteq \mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_3$ and thus

$$\mathbb{P}\{\mathcal{E}_0\} \leq \mathbb{P}\{\mathcal{E}_1\} + \mathbb{P}\{\mathcal{E}_2 | \mathcal{E}_1^c\} + \mathbb{P}\{\mathcal{E}_3 | \mathcal{E}_1^c\} \quad (236)$$

$$\stackrel{(a)}{\leq} \mathbb{P}\{\mathcal{E}_2|\mathcal{E}_1^c\} + \mathbb{P}\{\mathcal{E}_3|\mathcal{E}_1^c\} + \varepsilon', \quad (237)$$

where (a) follows from (231). We can partition $\widetilde{\mathfrak{M}}_{\mathcal{A}_a}$ into (random) subsets $\mathcal{D}_{\mathcal{A}'}$, indexed by $\mathcal{A}' \subseteq \mathcal{A}_a$ as

$$\mathcal{D}_{\mathcal{A}'} \triangleq \{\tilde{w}_{\mathcal{A}_a} \in \widetilde{\mathfrak{M}}_{\mathcal{A}_a} : \tilde{w}_{\mathcal{A}'^c} = \tilde{W}_{\mathcal{A}'^c} \wedge \tilde{w}_k \neq \tilde{W}_k, \forall k \in \mathcal{A}'\}, \quad (238)$$

where we used $\mathcal{A}'^c \triangleq \mathcal{A}_a \setminus \mathcal{A}'$. Observe that $\mathcal{D}_\emptyset = \{\tilde{W}_{\mathcal{A}_a}\}$. For each set $\emptyset \neq \mathcal{A}' \subseteq \mathcal{A}_a$ we define the error event

$$\mathcal{E}_{\mathcal{A}'} \triangleq \{\mathbf{V}_{\mathcal{D}_{\mathcal{A}'}}^{(\mathcal{A}_a)} \cap \mathcal{T}_{[\mathbf{U}_{\mathcal{A}_a}]\delta}^n \neq \emptyset\} \quad (239)$$

and we have

$$\mathcal{E}_2 \subseteq \bigcup_{\substack{\mathcal{A}' \subseteq \mathcal{A}_a : \\ \mathcal{A}' \cap \mathcal{A}_b \neq \emptyset}} \mathcal{E}_{\mathcal{A}'} \quad (240)$$

which implies

$$\mathbb{P}\{\mathcal{E}_2|\mathcal{E}_1^c\} \leq \sum_{\substack{\mathcal{A}' \subseteq \mathcal{A}_a : \\ \mathcal{A}' \cap \mathcal{A}_b \neq \emptyset}} \mathbb{P}\{\mathcal{E}_{\mathcal{A}'}|\mathcal{E}_1^c\}. \quad (241)$$

By the construction of the codebook, $\mathcal{D}_{\mathcal{A}'}$ has $\prod_{k \in \mathcal{A}'} (e^{n\Delta_k} - 1)$ elements. For $\tilde{w}_{\mathcal{A}_a} \in \mathcal{D}_{\mathcal{A}'}$, we have that $\mathbf{V}_{\tilde{w}_{\mathcal{A}'}}^{(\mathcal{A}')}$ are uniformly distributed on $\prod_{k \in \mathcal{A}'} \mathcal{T}_{[\mathbf{U}_k]\delta}^n$ and $\tilde{w}_{\mathcal{A}'^c} = \tilde{W}_{\mathcal{A}'^c}$. Given \mathcal{E}_1^c we have in particular $\tilde{\mathbf{U}}_{\mathcal{A}_a} \in \mathcal{T}_{[\mathbf{U}_{\mathcal{A}_a}]\delta}^n$. Thus, for any $\mathbf{u}_{\mathcal{A}'^c} \in \mathcal{T}_{[\mathbf{U}_{\mathcal{A}'^c}]\delta}^n$, we can conclude,

$$\mathbb{P}\{\mathcal{E}_{\mathcal{A}'}|\mathcal{E}_1^c, \tilde{\mathbf{U}}_{\mathcal{A}'^c} = \mathbf{u}_{\mathcal{A}'^c}\} = \mathbb{P}\left\{\bigcup_{\tilde{w}_{\mathcal{A}_a} \in \mathcal{D}_{\mathcal{A}'}} \{\mathbf{V}_{\tilde{w}_{\mathcal{A}'}}^{(\mathcal{A}_a)} \in \mathcal{T}_{[\mathbf{U}_{\mathcal{A}_a}]\delta}^n\} \middle| \mathcal{E}_1^c, \tilde{\mathbf{U}}_{\mathcal{A}'^c} = \mathbf{u}_{\mathcal{A}'^c}\right\} \quad (242)$$

$$\leq \sum_{\tilde{w}_{\mathcal{A}_a} \in \mathcal{D}_{\mathcal{A}'}} \mathbb{P}\{\mathbf{V}_{\tilde{w}_{\mathcal{A}'}}^{(\mathcal{A}_a)} \in \mathcal{T}_{[\mathbf{U}_{\mathcal{A}_a}]\delta}^n | \mathcal{E}_1^c, \tilde{\mathbf{U}}_{\mathcal{A}'^c} = \mathbf{u}_{\mathcal{A}'^c}\} \quad (243)$$

$$\leq \exp\left(n\left(\sum_{k \in \mathcal{A}'} \Delta_k\right)\right) \frac{|\mathcal{T}_{[\mathbf{U}_{\mathcal{A}'}|\mathbf{U}_{\mathcal{A}'^c}]\delta}^n(\mathbf{u}_{\mathcal{A}'^c})|}{\prod_{k \in \mathcal{A}'} |\mathcal{T}_{[\mathbf{U}_k]\delta}^n|} \quad (244)$$

$$\stackrel{(a)}{\leq} \exp\left(n\left(\sum_{k \in \mathcal{A}'} \Delta_k\right)\right) \frac{\exp(n(\mathbb{H}(\mathbf{U}_{\mathcal{A}'}|\mathbf{U}_{\mathcal{A}'^c}) + \varepsilon_0(\delta)))}{\exp(n(\sum_{k \in \mathcal{A}'} \mathbb{H}(\mathbf{U}_k) - \varepsilon_k(\delta)))} \quad (245)$$

$$\leq \exp\left(n\left(\varepsilon(\delta) + \mathbb{H}(\mathbf{U}_{\mathcal{A}'}|\mathbf{U}_{\mathcal{A}'^c}) + \sum_{k \in \mathcal{A}'} (\Delta_k - \mathbb{H}(\mathbf{U}_k))\right)\right), \quad (246)$$

where $\varepsilon(\delta) = \sum_{k \in \mathcal{A}' \cup 0} \varepsilon_k(\delta)$ goes to zero as $\delta \rightarrow 0$. Here, (a) follows from (113) and part 2 of Lemma 25. We observe that the definition of \tilde{R}_k and (66) imply for any $\emptyset \neq \mathcal{A}' \subseteq \mathcal{A}_a$ with $\mathcal{A}' \cap \mathcal{A}_b \neq \emptyset$ that

$$\sum_{k \in \mathcal{A}'} \Delta_k \leq -\frac{\varepsilon''}{2} - \mathbb{H}(\mathbf{U}_{\mathcal{A}'}|\mathbf{U}_{\mathcal{A}'^c}) + \sum_{k \in \mathcal{A}'} \mathbb{H}(\mathbf{U}_k). \quad (247)$$

Marginalize over $\tilde{\mathbf{U}}_{\mathcal{A}'^c}$ in (246) and use (247) to obtain

$$\mathbb{P}\{\mathcal{E}_{\mathcal{A}'}|\mathcal{E}_1^c\} \leq \exp\left(n\left(\varepsilon(\delta) - \frac{\varepsilon''}{2}\right)\right) \leq \varepsilon' \quad (248)$$

for n large enough and δ small enough. Applying the same arguments to $\mathbb{P}\{\mathcal{E}_3|\mathcal{E}_1^c\}$ and combining (237), (241) and (248), we have

$$\mathbb{P}\{\mathcal{E}_0\} \leq \varepsilon' + 2^{|\mathcal{A}_a|} \varepsilon' + 2^{|\mathcal{B}_a|} \varepsilon' \leq 2^K \varepsilon'. \quad (249)$$

For a set $\emptyset \neq \mathcal{A} \subseteq \mathcal{K}$, we next analyze the random quantity $\mathbf{L} \triangleq |\mathcal{C}_{\mathcal{A}} \cap \mathcal{T}_{[\mathbf{U}_{\mathcal{A}]}^n}|$. For n large enough, we have

$$\mathbb{E}[\mathbf{L}] \leq \sum_{\mathbf{V}_{\mathcal{A}} \in \mathcal{C}_{\mathcal{A}}} \mathbb{E}[\mathbf{1}_{\mathcal{T}_{[\mathbf{U}_{\mathcal{A}]}^n}(\mathbf{V}_{\mathcal{A}})}] \quad (250)$$

$$= \left(\prod_{k \in \mathcal{A}} \widetilde{M}_k \right) \mathbb{E}[\mathbf{1}_{\mathcal{T}_{[\mathbf{U}_{\mathcal{A}]}^n}(\mathbf{V}_{\mathcal{A}})}] \quad \text{for any } \mathbf{V}_{\mathcal{A}} \in \mathcal{C}_{\mathcal{A}} \quad (251)$$

$$= \left(\prod_{k \in \mathcal{A}} \widetilde{M}_k \right) \frac{|\mathcal{T}_{[\mathbf{U}_{\mathcal{A}]}^n}|}{\prod_{k \in \mathcal{A}} |\mathcal{T}_{[\mathbf{U}_k]^n}|} \quad (252)$$

$$\stackrel{(a)}{\leq} \left(\prod_{k \in \mathcal{A}} \widetilde{M}_k \right) \frac{e^{n(\mathbf{H}(\mathbf{U}_{\mathcal{A}}) + \varepsilon_0(\delta))}}{e^{n(\sum_{k \in \mathcal{A}} \mathbf{H}(\mathbf{U}_k) - \varepsilon_k(\delta))}} \quad (253)$$

$$\leq \left(\prod_{k \in \mathcal{A}} \widetilde{M}_k \right) e^{n(\mathbf{H}(\mathbf{U}_{\mathcal{A}}) - \sum_{k \in \mathcal{A}} \mathbf{H}(\mathbf{U}_k) + \hat{\varepsilon}(\delta))} \quad (254)$$

$$= \exp \left(n \left(\mathbf{H}(\mathbf{U}_{\mathcal{A}}) + \hat{\varepsilon}(\delta) + \sum_{k \in \mathcal{A}} \mathbf{I}(\mathbf{U}_k; \mathbf{X}_k) + \frac{\varepsilon''}{2} - \mathbf{H}(\mathbf{U}_k) \right) \right) \quad (255)$$

$$= \exp \left(n \left(\mathbf{H}(\mathbf{U}_{\mathcal{A}}) + \hat{\varepsilon}(\delta) + |\mathcal{A}| \frac{\varepsilon''}{2} - \sum_{k \in \mathcal{A}} \mathbf{H}(\mathbf{U}_k | \mathbf{X}_k) \right) \right) \quad (256)$$

$$= \exp \left(n \left(\mathbf{I}(\mathbf{U}_{\mathcal{A}}; \mathbf{X}_{\mathcal{A}}) + \hat{\varepsilon}(\delta) + |\mathcal{A}| \frac{\varepsilon''}{2} \right) \right). \quad (257)$$

where $\hat{\varepsilon}(\delta) = \sum_{k \in \mathcal{A} \cup 0} \varepsilon_k(\delta)$ goes to zero as $\delta \rightarrow 0$. Here, (a) follows from parts 1 and 2 of Lemma 25. We can choose ε'' so small that $\hat{\varepsilon}(\delta) + K\varepsilon''/2 < \varepsilon$ for sufficiently small δ . Defining the error event $\mathcal{E}_4 = \{\mathbf{L} \geq \exp(n(\mathbf{I}(\mathbf{U}_{\mathcal{A}}; \mathbf{X}_{\mathcal{A}}) + \varepsilon))\}$ we know from Markov's inequality that for n large enough

$$\mathbb{P}\{\mathcal{E}_4\} \leq \exp \left(n \left(\hat{\varepsilon}(\delta) - \varepsilon + |\mathcal{A}| \frac{\varepsilon''}{2} \right) \right) \leq \varepsilon'. \quad (258)$$

Using (249) and (258) we can apply Lemma 27 and obtain deterministic encoding functions $f_k: \mathcal{X}_k^n \rightarrow \mathcal{M}_k$, and deterministic decoding functions $g_{\mathcal{A}_a, \mathcal{A}_b}: \mathcal{M}_{\mathcal{A}_a} \rightarrow \mathcal{U}_{\mathcal{A}_b}^n$, such that (77) holds whenever the conditions of part 1 are satisfied. Taking into account that $g_{\mathcal{A}_a, \mathcal{A}_b}(\mathcal{M}_{\mathcal{A}_a}) \times g_{\mathcal{B}_a, \mathcal{B}_b}(\mathcal{M}_{\mathcal{B}_a}) \subseteq \mathcal{C}_{\mathcal{A}_b \cup \mathcal{B}_b}$, we also have (78). (Note that, given a specific code, $\mathbb{P}\{\mathcal{E}_4\} < 1$ already implies $\mathbb{P}\{\mathcal{E}_4\} = 0$ as the event \mathcal{E}_4 is fully determined by the code $\mathcal{C}_{\mathcal{K}}$ alone.)

J. A Random Coding Lemma

Lemma 27. *Let \mathbf{X} be discrete random variables and for any $\delta > 0$ let \mathbf{F}_{δ} be a random code (random vector-valued function) operating on \mathbf{X}^n (n sufficiently large as a function of δ). For finitely many error events $(\mathcal{E}_i)_{i \in \mathcal{I}}$ we have*

$$\mathbb{P}\{\mathcal{E}_i\} \leq \delta, \quad i \in \mathcal{I}. \quad (259)$$

Then, for any $\varepsilon > 0$ we can find $\delta > 0$, a sufficiently large $n \in \mathbb{N}$, and a code \mathbf{f} , such that

$$\mathbb{P}\{\mathcal{E}_i | \mathbf{F}_{\delta} = \mathbf{f}\} \leq \varepsilon, \quad i \in \mathcal{I}. \quad (260)$$

Proof: We can apply Markov's inequality to the random variable $\mathbb{P}\{\mathcal{E}_i | \mathbf{F}_{\delta}\}$ and obtain

$$\mathbb{P}\left\{ \mathbb{P}\{\mathcal{E}_i | \mathbf{F}_{\delta}\} \geq \sqrt{\delta} \right\} \leq \frac{\delta}{\sqrt{\delta}} = \sqrt{\delta}. \quad (261)$$

Applying the union bound yields

$$\mathbb{P}\left\{ \bigcup_{i \in \mathcal{I}} \left\{ \mathbb{P}\{\mathcal{E}_i | \mathbf{F}_{\delta}\} \geq \sqrt{\delta} \right\} \right\} \leq \sum_{i \in \mathcal{I}} \mathbb{P}\left\{ \mathbb{P}\{\mathcal{E}_i | \mathbf{F}_{\delta}\} \geq \sqrt{\delta} \right\} \quad (262)$$

$$\leq |\mathcal{I}|\sqrt{\delta}. \quad (263)$$

Thus, with a probability no smaller than $1 - |\mathcal{I}|\sqrt{\delta}$ the random coding yields a specific code \mathbf{f} such that $P\{\mathcal{E}_i|\mathbf{F}_\delta\} < \sqrt{\delta}$ for all $i \in \mathcal{I}$. Choosing $\delta = \min(\varepsilon^2, |\mathcal{I}|^{-2}/2)$ yields the desired result. ■

ACKNOWLEDGMENT

The authors would like to thank Shlomo Shamai (Shitz) and Emre Telatar for insightful discussions regarding the binary example.

REFERENCES

- [1] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," in *Claude Elwood Shannon: collected papers*, N. J. A. Sloane and A. D. Wyner, Eds. IEEE Press, 1993, pp. 325–350.
- [2] J. A. Hartigan, "Direct clustering of a data matrix," *Journal of the American Statistical Association*, vol. 67, no. 337, pp. 123–129, Mar. 1972.
- [3] B. Mirkin, *Mathematical Classification and Clustering*. Kluwer Academic Publisher, 1996.
- [4] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: a survey," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 1, no. 1, pp. 24–45, Jan. 2004.
- [5] Y. Cheng and G. M. Church, "Biclustering of expression data," in *Proc. 8th Int. Conf. Intelligent Syst. for Molecular Biology*, vol. 8, San Diego, CA, USA, Aug. 2000, pp. 93–103.
- [6] A. Tanay, R. Sharan, and R. Shamir, "Biclustering algorithms: A survey," *Handbook of Computational Molecular Biology*, vol. 9, no. 1-20, pp. 122–124, 2005.
- [7] R. Sharan, "Analysis of biological networks: Network modules – clustering and biclustering," lecture notes, 2006. [Online]. Available: <http://www.cs.tau.ac.il/~roded/courses/bnet07.html>
- [8] N. Slonim, G. S. Atwal, G. Tkačik, and W. Bialek, "Information-based clustering," *Proc. of the Nat. Academy of Sciences of the United States of America*, vol. 102, no. 51, pp. 18 297–18 302, 2005.
- [9] I. S. Dhillon, S. Mallela, and D. S. Modha, "Information-theoretic co-clustering," in *Proc. of the 9th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Washington, DC, USA, Aug. 2003, pp. 89–98.
- [10] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge University Press, 2011.
- [11] T. S. Han, "Hypothesis testing with multiterminal data compression," *IEEE Trans. Inf. Theory*, vol. 33, no. 6, pp. 759–772, Nov. 1987.
- [12] T. S. Han and S. Amari, "Statistical inference under multiterminal data compression," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2300–2324, Oct. 1998.
- [13] R. Ahlswede and I. Csiszár, "Hypothesis testing with communication constraints," *IEEE Trans. Inf. Theory*, vol. 32, no. 4, pp. 533–542, Jul. 1986.
- [14] J. Körner and K. Marton, "How to encode the modulo-two sum of binary sources," *IEEE Trans. Inf. Theory*, vol. 25, no. 2, pp. 219–221, Mar. 1979.
- [15] C. Nair, "Upper concave envelopes and auxiliary random variables," *Int. J. of Advances in Eng. Sciences and Appl. Math.*, vol. 5, no. 1, pp. 12–20, 2013.
- [16] T. A. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," *IEEE Trans. Inf. Theory*, vol. 60, no. 1, pp. 740–761, Jan. 2014.
- [17] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. of the 37th Annu. Allerton Conf. on Commun., Control, and Computing*, Monticello, IL, Sep. 1999, pp. 368–377.
- [18] R. Gilad-Bachrach, A. Navot, and N. Tishby, "An information theoretic tradeoff between complexity and accuracy," in *Learning Theory and Kernel Machines*. Springer, 2003, pp. 595–609.
- [19] R. Ahlswede and J. Körner, "Source coding with side information and a converse for degraded broadcast channels," *IEEE Trans. Inf. Theory*, vol. 21, no. 6, pp. 629–637, Nov. 1975.
- [20] A. D. Wyner, "On source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. 21, no. 3, pp. 294–300, May 1975.
- [21] H. S. Witsenhausen and A. D. Wyner, "A conditional entropy bound for a pair of discrete random variables," *IEEE Trans. Inf. Theory*, vol. 21, no. 5, pp. 493–501, Sep. 1975.
- [22] A. Wyner and J. Ziv, "A theorem on the entropy of certain binary sequences and applications: Part I," *IEEE Trans. Inf. Theory*, vol. 19, no. 6, pp. 769–772, Nov. 1973.
- [23] H. Witsenhausen, "Entropy inequalities for discrete channels," *IEEE Trans. Inf. Theory*, vol. 20, no. 5, pp. 610–616, Sep. 1974.
- [24] A. Wyner, "A theorem on the entropy of certain binary sequences and applications: Part II," *IEEE Trans. Inf. Theory*, vol. 19, no. 6, pp. 772–777, Nov. 1973.
- [25] R. Ahlswede and J. Körner, "On the connection between the entropies of input and output distributions of discrete memoryless channels," in *Proc. 5th Conf. Probability Theory, Sep. 1974*, Brasov, Romania, 1977, pp. 13–23.
- [26] E. Erkip and T. M. Cover, "The efficiency of investment information," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1026–1040, May 1998.
- [27] G. R. Kumar and T. A. Courtade, "Which Boolean functions are most informative?" in *Proc. IEEE Int. Symp. on Inform. Theory*, Istanbul, Turkey, Jul. 2013, pp. 226–230.
- [28] T. A. Courtade and G. R. Kumar, "Which Boolean functions maximize mutual information on noisy inputs?" *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4515–4525, Aug. 2014.

- [29] J. G. Klotz, D. Kracht, M. Bossert, and S. Schober, "Canalizing Boolean functions maximize mutual information," *IEEE Trans. Inf. Theory*, vol. 60, no. 4, pp. 2139–2147, Apr. 2014.
- [30] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2006.
- [31] A. Orlitsky and J. R. Roche, "Coding for computing," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 903–917, Mar. 2001.
- [32] S.-Y. Tung, "Multiterminal source coding," Ph.D. dissertation, Cornell University, May 1978.
- [33] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011.
- [34] P. Gács and J. Körner, "Common information is far less than mutual information," *Problems of Control and Inform. Theory*, vol. 2, pp. 149–162, 1973.
- [35] T. Berger, "Multiterminal source coding," in *The Information Theory Approach to Communications*, G. Longo, Ed. Springer, 1977, vol. 229, pp. 171–231.
- [36] T. Berger, Z. Zhang, and H. Viswanathan, "The CEO problem," *IEEE Trans. Inf. Theory*, vol. 42, no. 3, pp. 887–902, May 1996.
- [37] T. S. Han and K. Kobayashi, "A unified achievable rate region for a general class of multiterminal source coding systems," *IEEE Trans. Inf. Theory*, vol. 26, no. 3, pp. 277–288, May 1980.
- [38] B. Grünbaum, *Convex Polytopes*. Springer, New York, 2003.
- [39] A. A. Gohari and V. Anantharam, "Evaluation of Marton's inner bound for the general broadcast channel," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 608–619, Feb. 2012.
- [40] V. Jog and C. Nair, "An information inequality for the BSSC broadcast channel," in *Inform. Theory and Applicat. Workshop (ITA)*, San Diego, CA, USA, Feb. 2010, pp. 1–8.
- [41] R. Schneider, *Convex Bodies: The Brunn-Minkowski Theory*, 2nd ed. Cambridge University Press, 2014.
- [42] W. Rudin, *Functional Analysis*, 2nd ed. McGraw-Hill, 1991.
- [43] H. G. Eggleston, *Convexity*, P. Hall and F. Smithies, Eds. Cambridge University Press, 1958.
- [44] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. 22, no. 1, pp. 1–10, Jan. 1976.
- [45] J. R. Munkres, *Topology*. Prentice Hall, 2000.
- [46] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. McGraw-Hill, 1976.
- [47] C. D. Aliprantis and K. C. Border, *Infinite Dimensional Analysis: A Hitchhiker's Guide*, 3rd ed. Springer, 2006.